Catalogue no. 71-526-XPB

Methodology of the Canadian Labour Force Survey

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this product and related statistics or services should be directed to: Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-9809) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialling area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no. 71-526-XPB, is published occasionally as a standard paper product for $50.00 in Canada. Outside Canada the cost is US$50.00.

Please order by mail, at Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, K1A 0T6; by phone, at **(613) 951-7277** or **1 800 770-1033**; by fax, at **(613) 951-1584** or **1 800 889-9734**; or by Internet, at order@statcan.ca. For changes of address, please provide both old and new addresses. Statistics Canada products may also be purchased from authorized agents, bookstores and local Statistics Canada offices.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

ISBN 0-660-60566-X

9 780660 605661

71-526-XPB 98001

Statistics Canada
Household Survey Methods Division

Methodology of the Canadian Labour Force Survey

J.G. Gambino, M.P. Singh, J. Dufour, B. Kennedy, J. Lindeyer

Published by authority of the Minister responsible for Statistics Canada

December 1998

Catalogue no. 71-526-XPB

Frequency: Occasional

ISBN 0-660-60566-X

Ottawa

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Acknowledgement

Many groups within Statistics Canada are involved in the design and operations of the Labour Force Survey. Divisions or Branches with a major involvement in the survey include the following: Household Surveys Division, responsible for management of the survey, data dissemination and user liaison; Survey Operations Branch, responsible for field operations, and data capture and processing in Regional Offices; Methodology Branch, responsible for the sample design, data collection methodology, estimation procedures, and quality evaluation; and Informatics Branch, responsible for electronic data processing services to both Ottawa and the Regional Offices.

In 1990, a Redesign Steering Committee was established with members from the above areas, along with members from other areas whose programs are associated with the Labour Force Survey infrastructure. The Committee met until 1996 to provide direction and guidance on the development and implementation of the sample redesign.

The major contributors to this report were Jack Gambino, M.P. Singh, Johane Dufour, Brian Kennedy and John Lindeyer. Others contributors and commentators included Doug Drew, Mike Sheridan, Deborah Sunter and Diane Stukel.

Table of Contents

CHAPTER 1: Introduction

Background

The Canadian Labour Force Survey (LFS) was introduced following the Second World War to satisfy a need for reliable and timely data on the labour market. Information was urgently required on the massive labour market changes during the transition from a war-time to a peace-time economy. The survey was designed to provide estimates of employment and unemployment at the regional and national level.

The LFS began as a quarterly survey in 1945 and became a monthly survey in 1952. In 1960, the Interdepartmental Committee on Unemployment Statistics recommended that the LFS be designated the source of the official measure of unemployment in Canada. This endorsement was followed by a demand for a broader range of labour market statistics, including more detailed regional data. The information generated by the survey has expanded considerably over the years and now provides a detailed picture of the Canadian labour market.

Concepts and Outputs

The LFS is the only official source of monthly estimates of total employment (both paid workers and the self-employed, as well as full-time and part-time workers) and unemployment. Key rates published monthly include the unemployment rate (defined as the number of unemployed people divided by the size of the labour force), the employment rate (defined as the number of employed people divided by the total population, i.e., the number of people aged 15 years and older), and the participation rate (the proportion of the population that is either employed or unemployed). It is a major source of information on the personal characteristics of the working-age population, including age, marital status, educational attainment and family characteristics.

Employment estimates include detailed breakdowns by industry and occupation, job tenure, and usual and actual hours worked. The survey incorporates questions permitting analyses of many topical issues, such as involuntary part-time employment, multiple job-holding, and absence from work.

Unemployment estimates are produced by industry and occupation as well as by duration of unemployment, type of work sought, and activity before looking for work. Information is also available on the recent labour market involvement of persons not in the labour force. For a full description of the content of the LFS questionnaire, see Statistics Canada (1998)[1].

As part of the redesign of the LFS, a new questionnaire was introduced effective January 1997. The new questionnaire adds new content, including wages and union membership. Sunter et al. (1995) give a detailed description of the process of redesigning the LFS questionnaire.

In addition to providing national and provincial estimates, the Labour Force Survey also releases estimates for subprovincial areas such as Employment Insurance Economic Regions (EIERs) and Census Metropolitan Areas (CMAs). In recent years, estimates of the standard labour market indicators have also been tabulated for small areas, such as Census Divisions (CDs) and Canada Employment Centres, using special estimation techniques. LFS estimates are used by both federal and provincial governments in allocating funds and other resources among various political and administrative jurisdictions.

Standard estimates from the LFS are published monthly in *Labour Force Information* (Catalogue 71-001-PPB). A variety of labour market information is also available from CANSIM, Statistics Canada's electronic database and retrieval system. Over nine thousand time series on this database are updated monthly by the LFS.

Beginning in 1997, the quarterly *Labour Force Update* (Catalogue 71-005-XPB) provides an in-depth look at a variety of subjects central to analysis of the labour market. Each issue has a specific focus, and topics such as hours of work, youths in the labour market and wages will be assessed on a regular basis.

The LFS also produces an annual *Labour Force Historical Review on CD-ROM* (Catalogue 71F0004XCB) containing comprehensive data in both cross-sectional and time series format from 1976 to the current year.

---

[1] References are given before Appendix A.

A great deal more information is available from the survey than can be published regularly. Requests for custom tabulations are filled on a cost-recovery basis.

An Overview of the Survey

Target population. The LFS covers 98 percent of the Canadian population. The survey excludes from its coverage the Northwest Territories and residents of Indian reserves and Crown lands. Also excluded are inmates of institutions and full-time members of the Canadian Armed Forces since both groups are considered to be outside the labour market. The survey establishes the labour force status of all members of selected households who are 15 years of age and older.

Sample size. At the time of this writing, the target sample size for the LFS is 52,350 households. However, it has varied over time. During the 1970s redesign, the size of the sample was increased from 35,000 households per month to 55,000 households to meet the increasing demand for more reliable and detailed data at the provincial level. The LFS has also occasionally experienced reductions in its monthly samples. Following two such reductions in the 1980s, the sample comprised about 47,000 households. In 1989, the sample was increased by 16,500 households to 63,000 households. The aim of the increase was to produce better estimates for Unemployment Insurance Regions. The sample was decreased to about 59,000 households in 1993. Based on the efficiency gains from the new design, the sample was decreased further to 52,350 households, effective July 1995. With legislative changes introduced in June 1996, the boundaries of the UI regions were revised and the regions were renamed Employment Insurance Economic Regions (EIERs).

Sample rotation. The LFS follows a rotating panel sample design, in which households remain in the sample for six consecutive months. The sample is split into six representative sub-samples, and each month one-sixth of the sample is replaced after it has completed its stay in the survey. This results in a five-sixths month-to-month sample overlap, which makes the design efficient for estimating month-to-month changes. The rotation after six months prevents undue respondent burden for households that are selected for the survey.

Data collection. The LFS reference week is normally the week containing the fifteenth day of the month. Data collection for the LFS is carried out during the week following reference week. Statistics Canada employs about 850 interviewers, including 80 senior interviewers, across the country, and their data collection activities are managed from five Regional Offices (ROs). LFS interviews are conducted in person in the first month and by telephone in the five subsequent months. The questionnaires are completed by the interviewer using Computer Assisted Interviewing (CAI) on portable computers. The collected data are uploaded daily by the interviewer to the RO during the interview week and then transmitted to Ottawa for processing. The data collection, data processing and dissemination processes are streamlined and efficiently run. As a result, Statistics Canada publishes monthly LFS estimates just 13 days after the end of the interview week.

Due to the importance of the statistics produced from the survey and the complexities of the operations involved, various quality evaluation and control programs designed for different phases of the survey are monitored regularly.

Survey redesign. Following each decennial population census, the LFS has undergone a sample redesign to reflect changes in population characteristics and to respond to changes in the information needs to be satisfied by the survey. The redesign that took place following the 1971 census was the biggest redesign prior to the current one. Not only was the sample design modified, but also, substantial changes were made to the questionnaire, and a new data processing infrastructure was introduced. The redesign that took place following the 1981 census focused mainly on updating the sample design itself. The current redesign is again a major one, encompassing all aspects of the survey: computer assisted interviewing has been introduced, the sample design has changed, data processing and dissemination systems have undergone a complete overhaul, and the questionnaire has been revised substantially.

Objectives of the Sample Redesign

The current sample redesign program culminated with the introduction of a new sample at the end of 1994. The program included extensive consultation involving the reassessment not only of the survey's principal role as a provider of current labour market information, but also of its use as a central vehicle for conducting household surveys within Statistics Canada.

A redesign provides an opportunity to update the sampling frame, its stratification and the allocation of the sample to reflect changes in population size and distribution. The design in place until the end of 1994

used 1981 census geography and the corresponding 1981 census counts to select the sample and derive estimation weights. Since many standard geographical units change with every census, each redesign enables the LFS to adopt the most recent definitions of these units.

A goal in previous redesigns had been to make the LFS frame, sample and systems more flexible for other surveys since many household surveys conducted by Statistics Canada use these to meet their own needs. This goal was equally important in the current redesign. A related goal common to the last two redesigns was to take advantage of changes in technology and field operations to simplify the design.

A Note on UI Regions and EIE Regions. It was noted above in the overview that in 1995, Unemployment Insurance (UI) Regions were replaced by Employment Insurance Economic Regions, or EIERs. For convenience, we will use the new name throughout this document even though the sample redesign was based on the UI Regions. The supplementary sample of 16,500 households that was introduced in 1989 to improve estimates for EIERs will be referred to as the *EI sample*. The remaining sample, which currently consists of 35,850 households, will be referred to as the *core sample*.

A new goal for this redesign was to use the core sample to satisfy the requirements for national and provincial estimates while using the EI sample to optimize estimates for the EIE regions. How this was achieved will be explained in chapter 2.

Scope of the Report

This report is a reference on the methodological aspects of the Labour Force Survey. The survey design, estimation methodology and data quality are discussed in detail, with supplementary references given where appropriate. This document is complemented by a separate report entitled *Guide to the Labour Force Survey* (available on the internet at www.statcan.ca/english/concepts/labour/index.htm) which emphasizes subject matter issues and the data that is available from the survey.

There are five additional chapters in this report. Chapter 2 describes stratification and the formation of sampling units such as clusters. Chapter 3 gives a detailed discussion of sample allocation, sample selection in urban and rural areas and rotation of the sample over time. The use of the LFS sample and frame by other household surveys is discussed briefly in Chapter 4.

Chapter 5 presents a detailed description of the Labour Force Survey's weighting and estimation system, including the handling of nonresponse. Finally, the LFS has an extensive data quality monitoring program which is described in Chapter 6.

Note: A list of abbreviations used in this document is given in Appendix B. A diagram summarizing the new Labour Force Survey design is given in Appendix C.

CHAPTER 2: Stratification and Sampling Unit Formation

Canada's population lives in various geographical areas such as provinces and regions which have standard definitions. Survey samplers usually partition the population further into strata, from which samples are then selected independently. If the population units in each stratum are relatively homogeneous, then the size of the sample required to obtain estimates of a specified precision will be much smaller than the size needed under an unstratified design. Stratification has other useful properties: different designs, sampling schemes and estimation methods can be used in different strata, selective updating of the design only in strata that have undergone rapid change can be done, and strata can sometimes be useful operational units. In this section, we describe the stratification used by the Labour Force Survey.

Most surveys use two types of strata: (i) existing standard geographical units such as metropolitan areas and (ii) strata formed by combining smaller units such as census enumeration areas according to an objective criterion. The standard geographical units used as strata by the LFS are described first.

With the exception of Prince Edward Island, each province is divided into Economic Regions (ERs). The LFS has used ERs as primary strata since the 1960s. In the early stages of the current redesign, the definitions of the ERs were reviewed in consultation with the provinces. There are now 72 ERs in Canada.

In previous LFS designs, ERs were the only subprovincial regions taken into account when designing the survey. In 1989, Human Resources Development Canada (HRDC) started funding an increase in the LFS sample of 16,500 households per month. This sample is used to improve the labour force estimates produced for the former 61 (now 53) Employment Insurance Economic Regions (EIERs). Thus, in the new design, both ERs and EIERs were taken into account in stratification. In allocation, relatively less attention was paid to improving estimates for ERs since the core sample was allocated primarily with provincial and national estimates in mind, while the extra sample funded by HRDC was directly targeted to EIE regions.

The two sets of regions, while roughly equal in number, were defined for different purposes and do not usually coincide. To deal with both sets of regions in the new LFS design, the intersections of the regions were considered as basic strata. Due to a degree of overlap between the EIERs and ERs, there were 133 such intersections.

A third set of regions, the Census Metropolitan Areas, has also been respected by the LFS stratification in the current and previous designs. CMAs are urban areas whose population was at least 100,000 in the most recent census. All CMAs are also EIERs.

In previous designs, by the time a new design was in place, the CMA definitions it used were already four years old. For example, the previous design was fully in place in March 1985 and used CMA definitions from the June 1981 census. In the new design, the work to define CMAs for the 1996 Census was started early to accommodate the LFS redesign which is based on these new CMA definitions. The official 1996 CMAs differ from the ones used initially by the LFS since legal changes in municipal boundaries must be taken into account. The differences are minor, and adjustments are made to the LFS to respect the final CMA definitions. The 1996 CMAs have also been adopted as EIERs.

Within the larger geographical strata, more detailed strata were formed without regard to geographical constraints. This was done using the same method used in the previous design, namely a clustering algorithm due to Friedman and Rubin (1967) and modified for the LFS by Drew et al. (1985). The algorithm partitions units into strata that are as homogeneous as possible for several variables by minimizing a within-group weighted sum of squares. The sums of squares are computed to reflect the sampling of units with probability proportional to size. Different weights can be assigned to different variables if desired. More details of the algorithm can be found in Drew et al. (1985) and in Singh et al. (1990). Stratification using this algorithm will be referred to as *optimal stratification*.

Stratification variables. The variables used in the stratification program, given below, include all those used in the previous LFS design. However, the new design uses a more detailed breakdown of employment by industry, specifically for the manufacturing and services sectors. The only completely new stratification variables

are based on mother tongue. For each stratification unit, three language variables were coded: the number of people who gave English, French or Other (i.e., any other language) as their mother tongue. Finally, the income variable was given three times the weight of the other stratification variables.

Data from the 1991 census were used for stratification. The following variables were used.

Employed in
- agriculture
- forestry and fishing
- mining
- manufacturing - consumables
- manufacturing - rubber, plastics, leather
- manufacturing - textiles and clothing
- manufacturing - furniture, pulp and paper, printing, wood
- manufacturing - metals and minerals
- manufacturing - petrochemical, chemical
- construction
- transportation
- services - trade
- services - financial
- services - personal/business
- services - government

total employed

total income
population aged 15+
population aged 15-24
population aged 55+
number of one-person households
number of two-person households
number of owned dwellings
total gross rent
population with high school education
mother tongue English
mother tongue French
mother tongue other than English/French

The choice of stratification variables was customized for each area that underwent optimal stratification. Within areas being optimally stratified, the above variables were obtained from the 1991 census. If a variable accounted for less than two percent of the total population, then it was dropped. For groups such as services, if subgroups such as financial services were not significant, then the grouped variable was used instead. A group was considered significant if it accounted for more than two percent of the population.

Types of area for stratification. The LFS frame may be divided into three types of areas: (1) rural areas, (2) larger cities, (3) smaller urban areas. For stratification, each of these could be further divided, as described next.

1. Rural areas. In rural areas, strata were usually formed by manually grouping two or three Census Divisions within an ER-EIER intersection into a geographical stratum. Decisions regarding the formation of these geographical strata were made in conjunction with decisions regarding the choice of first-stage sampling units that would be most appropriate (e.g., two-stage EA versus three-stage PSU design) and whether there should be separate rural and urban strata formed. Optimal stratification within the geographical strata was performed whenever there was sufficient population to warrant it. In general, the rural strata in the new design tend to be smaller than the rural strata in the previous design.

2. Larger cities (population of 50,000 or more): In 17 CMAs, there were enough apartment buildings to form a separate list frame, referred to as the apartment frame, described in 2.3 below. Excluding the apartment frame, the remainder of each large urban centre comprised an area frame. In addition, where feasible, areas with high average income also formed a separate stratum, described in 2.2. The remaining dwellings formed the regular strata and are described below under *Street Network File (SNF) areas*. As will be noted below in (2.1), SNF areas correspond to CADP (computer assisted districting program) areas.

When more than one level of stratification exists in a city, we will refer to the lowest-level, smallest strata as final strata. Final strata were designed to have an expected sample size of at least 48 households (36 households in Toronto, Montreal and Vancouver). Expected sample sizes in Toronto, Montreal and Vancouver are smaller because in the new design, the sample yield per cluster was chosen to be about double the yield in the old design, and the latter yield was somewhat lower for these three cities.

2.1 Larger Cities: SNF (or CADP) areas. These urban areas are covered by Geography Division's Street Network File. They include the 25 CMAs and 20 of the largest Census Agglomerations (CAs). In these areas, there are up to three levels of stratification. Within a CMA or CA, if the anticipated sample size in a municipality (i.e., a Census Subdivision or CSD) within it was at least 240 households (180 households in Toronto, Montreal and Vancouver), then the municipality itself is a stratum. If the municipality is large enough to

form more than five final strata, it is stratified optimally into groups, called superstrata, that will in turn yield three (sometimes four or five) final strata. Superstrata are formed using Census Tracts (CTs) as stratification units (or CSDs in non-tracted outskirts of cities). These superstrata are geographically compact and contiguous. In Toronto, Montreal and Vancouver, the aim was to create superstrata that would yield six final strata instead of three.

If a CSD was not large enough to form more than five final strata, it was pooled with other such CSDs. This pool was then treated as a superstratum as in the previous paragraph, i.e., if large enough, optimal strata were formed within it, and so on.

Superstrata were divided optimally into the final strata; these are non-compact, non-contiguous strata yielding a sample of 48 households, except in Toronto, Montreal and Vancouver where they yield 36 households.

2.2 Larger Cities: High income strata. For the first time, high income strata were formed in the nine cities where this was feasible, namely, in Montreal, Ottawa, Toronto, Hamilton, London, Winnipeg, Edmonton, Calgary and Vancouver. In each of these cities, the three percent of Enumeration Areas (EAs) with the highest average household income in the 1991 census formed the high income stratum. Each stratum had to be large enough to yield a sample of 24 households. In five cities there were enough EAs to form two or more high income strata. This is summarized in Table A1 in Appendix A. In cities not on this list, there were not enough EAs with a high average household income (about $100,000) to form a separate stratum. More details on high income strata are given by Chen et al. (1994).

The introduction of high income strata is expected to make the representation of high income households in the sample more stable over time. This will benefit surveys, such as the Survey of Consumer Finances, that are based on the LFS frame or sample and collect income-related data. It will also help in the collection of earnings information in the new LFS questionnaire. Another possible benefit is that it will be easier to investigate whether the propensity to be nonrespondents is different for high income households. If it is established that nonresponse is substantially higher in high income strata, special measures to address the problem can be developed.

2.3 Larger Cities: The apartment frame. The LFS has maintained a list of apartment buildings in large CMAs since the 1960s. Currently, this list is used as a sampling frame in eighteen cities: Halifax, Quebec City, Montreal, Hull, Ottawa, Oshawa, Toronto, Hamilton, St. Catharines, Kitchener, London, Windsor, Winnipeg, Saskatoon, Calgary, Edmonton, Vancouver and Victoria. For the LFS, a building is treated as an apartment if it has at least five floors of living quarters and at least thirty residential units. In each city, as new buildings are constructed, they are added to the bottom of the list for that city. Since sampling of apartments is systematic, a new apartment building has a chance of falling in the sample as soon as it is built.

A new feature in the LFS design is the formation of a frame of low income apartment buildings. In contrast to the high income strata, it was found more beneficial to use apartment buildings instead of EAs for stratification of low income households.

An apartment building was added to the low income frame if its average household income according to the 1991 census was less than $20,000. For a low income frame to exist in a city, the frame must contain enough dwellings to yield a sample of at least 30 dwellings and the average income for the whole frame should be about $15,000. Low income frames were created in seven cities: Montreal, Ottawa (excluding Hull), Toronto, Winnipeg, Calgary, Edmonton and Vancouver.

Special cases: In Calgary and Edmonton, buildings with average incomes greater than $20,000 were added to the frame to bring the sample yield to the desired level. In Montreal, the low income frame exists entirely within ER 40.

Unlike the rest of the apartment frame, the low income frame is not open-ended since the average income of the residents of a newly-constructed apartment is not available.

Of the seven cities with a low income frame, only Toronto has enough units to support stratification of the frame. The CSD of Toronto and the rest of the CMA of Toronto each have a low income apartment frame.

For the non-low income apartments, an attempt was made to create list frames within geographic superstrata. This was possible in Halifax (2 apartment strata), Quebec (2), Montreal (4), Ottawa (3), Toronto (6), Hamilton (2), Kitchener (2) and Vancouver (4). These amount to geographical breakdowns of the overall apartment strata within these cities.

Finally, an attempt was made to further subdivide apartment strata according to apartment size. In each stratum, apartment buildings were classified according to size: less than 100 units, between 100 and 199 units, and 200 units or more. If there were enough apartments in a size class to yield a sample of thirty dwellings, it became a separate stratum. Otherwise it was collapsed with another class.

The apartment frame stratification is summarized in Table 1.

Table 1. Apartment Frame Strata

CMA	Geographical Strata	Total Number of Strata
Halifax	2	2
Quebec	2	2
Montreal*	4	9
Ottawa-Hull*	3	6
Oshawa	1	2
Toronto*	6	16
Hamilton	2	4
St. Catharines	1	1
Kitchener	2	2
London	1	2
Windsor	1	2
Winnipeg*	1	6
Saskatoon	1	1
Calgary*	1	3
Edmonton*	1	3
Vancouver*	4	6
Victoria	1	1
TOTAL	34	68

Note: (i) A * denotes that the city has low income strata.
(ii) The total number of strata includes low income strata.

3.1 Smaller urban areas: EA design (see Appendix D). In all cities except the smallest ones, non-compact, non-contiguous optimal final strata were formed using EAs as stratification units. In Sydney, Nova Scotia, first compact, contiguous superstrata were created, and then the final strata were formed within each superstratum.

3.2 Smaller urban areas: VR design (see Appendix D). In the smallest urban areas, if the area was classified as "self-representing" in the old design (i.e., it constituted at least one urban stratum with a sample size of at least 50 dwellings), its old stratification was used in the new design. In some cases, additions to the urban area were assigned manually to a stratum. The term *VR design* is used here because at the time of the previous redesign, counts from census visitation records (VRs) were used to form clusters.

Clustering in final strata. To reduce field costs, households in final strata are not selected directly. Instead each stratum is divided into clusters, and then a sample of clusters is selected in the stratum. Then, in each selected cluster, a sample of households is chosen. The methods used to select clusters and households are discussed later in the section on sample selection.

In rural areas, EAs are usually used as clusters. In urban areas, a variety of clusters are used. In the smallest urban areas, where the stratification from the previous design was usually retained, the old clusters were also used. These were formed manually using census visitation records: blockfaces were combined until the desired cluster size was achieved. The population counts for these clusters were updated using 1991 census information, and in some cases, the old clusters had to be modified because there had been major changes since the old design. In urban areas where stratification was based on EAs, EAs were also used as clusters. In some cases, large EAs were split into two clusters.

In the largest urban areas, i.e., those with a SNF, a new automated clustering method was used, replacing the labour-intensive manual operation used in the past. For the new design, the CADP program that was used to form EAs for the 1991 Census was modified by staff in Statistics Canada's Geography Division to form LFS clusters. The program combines block faces to produce clusters containing 150 to 200 dwellings (200 to 250 dwellings in Montreal, Toronto and Vancouver), on average.

The new urban clusters are about three times the size of clusters used in the previous design. The increased cluster size will help attenuate problems due to the rapid

population growth that occurs at times in urban areas, since the relative impact of growth will tend to be smaller for large clusters. The larger cluster size also reduces the frequency of cluster rotation.

Table 2 gives a summary of the types of first stage units used for the bulk of the LFS sample. The size refers to the number of households in a typical unit and can vary widely for a specific type of unit. The yield is the number of households selected by the LFS for interviewing in any given month.

Table 2. Major first-stage units, sizes and yields.

Area	Sampling Unit	Size (households per unit)	Yield (sampled households)
Toronto, Montreal, Vancouver	cluster	200-250	6
Other cities	cluster	150-200	8
Apartment frame	apartment	varies	5
Most rural areas and non-SNF parts of cities	EA	300	10

Appendix D lists all urban areas in Canada that are LFS strata or groups of strata, i.e., it lists all urban areas that always contain some LFS sample. This corresponds to the old self-representing unit concept in previous LFS designs. Any urban areas not listed in Appendix D are not strata. They are either primary sampling units, or parts of a larger urban stratum, or are incorporated into a rural stratum. Appendix D also indicates the types of clusters used in each area.

CHAPTER 3: Sample Allocation, Selection and Rotation

During the redesign, the total size of the monthly Labour Force Survey sample was kept at the level that existed in the old design, namely, 58,850 households. As part of the redesign process, this sample was allocated to best meet the need for good estimates at various geographical levels. These include the national, and provincial levels, Census Metropolitan Areas, and for the first time, Unemployment Insurance regions, which are now called Employment Insurance Economic Regions. The following were the reliability targets used.

Canada and the provinces: Maintain or improve CV levels for *unemployed* from the old design, i.e., a CV of about 2 percent for Canada and 4 to 7 percent for provinces.

EIERs/CMAs: CVs of 15 percent or less for quarterly estimates of *unemployed*. A minimum sample size per EIER of 600 households per month was set.

Though not a formal requirement, a target CV of 25 percent or less for quarterly estimates of *unemployed* for ERs was also used, although some collapsing of regions was necessary. There are 72 ERs, but for allocation, this number was reduced to 68 by collapsing a pair of ERs in each of Quebec, Manitoba, Saskatchewan and British Columbia. The CV target is for the collapsed regions. Table 3 summarizes the number of subprovincial regions in each province at the time of the redesign.

Table 3. ERs, EIERs and CMAs by province.

Province	ERs	EIERs	CMAs
Newfoundland	4	3	1
Prince Edward I.	1	1	0
Nova Scotia	5	5	1
New Brunswick	5	4	1
Quebec	16	13	6*
Ontario	11	18	10*
Manitoba	8	3	1
Saskatchewan	6	4	2
Alberta	8	4	2
British Columbia	8	6	2
Canada	72	61	25*

*The Ottawa-Hull CMA is counted in both Ontario and Quebec.

As was noted in the stratification section, ER-EIER intersections were used as strata. As a result, they were also the basic areas used during sample allocation. Because both ERs and EIERs tend to use census divisions as building blocks, there are only 133 intersections throughout Canada.

Allocation of the sample to provinces and to regions within provinces was discussed with the provinces as well as major users of LFS data. The final allocation was also affected by operational constraints. Several allocation strategies were studied, including Neyman, Kish, proportional, power and square root allocation. They are summarized by Mian and Laniel (1994). Here, only the approach actually implemented is described.

At the time of the redesign, the total LFS sample size consisted of a core sample of 42,310 households and a HRDC-funded sample of 16,500 households. Because of changes in the sample size and in the population during the years since the previous redesign, the allocation of the core sample was optimal for neither provincial nor subprovincial estimates.

The overall strategy was to first allocate the core sample to optimize provincial and national estimates. This was followed by allocation of the HRDC sample to supplement the core sample in the EIERs that need it most (typically, these will be regions with relatively small populations).

Allocation of core sample to provinces. With the following exceptions, the core provincial sample sizes that existed before the HRDC-funded increase were retained. The exceptions are: a transfer of sample from Saskatchewan to Manitoba and from Alberta to BC. The amount of sample transferred was just enough to give each pair of provinces equal CVs for *unemployed* based on the core sample. This was done to correct imbalances in the old allocation that were introduced via sample decreases that occurred in the life of the old design (provinces with larger CMAs were harder hit by decreases in sample since the cuts tended to be concentrated in large CMAs).

Allocation of core sample to ERs within provinces. For each province, the core sample was allocated to ERs in proportion to the size of these regions, where size is measured as the number of private occupied dwellings in the ER according to the 1991 census. This is an allocation primarily aimed at optimizing estimates at the provincial level. However, since sparsely populated ERs

receive too little sample if proportional allocation is followed strictly, a minimum sample size was set, namely 200 households per ER. In Alberta, where the minimum was set at 300 households, some efficiency at the provincial level was sacrificed to benefit small ERs.

Allocation within ERs. The samples allocated in the previous step were then allocated proportionally within each ER to ER-EIER intersections. Again, this is a nearly optimal allocation of the sample.

Allocation of the HRDC-funded sample to EIERs. The previous step completed the allocation of the core sample. The next step was to allocate the HRDC sample of 16,500 households. Since HRDC funds these extra households to ensure that labour force estimates for EIERs are of adequate quality, the sample was allocated to EIERs to maximize the improvement in CVs of *unemployed,* targeting the sample to EIERs that had high CVs based on the core sample alone. Allocation of the sample in this fashion resulted in a CV of 10 percent or better for the estimate of *unemployed* in each EIER. A minimum sample size of 600 households was allocated to each EIER.

Table A2 in Appendix A gives the core and total allocations by province for the old and the new design at the time of the redesign, as well as the allocation after the sample reduction that took place in 1995.

Calculation of CVs. At various stages in the allocation process, CVs for the characteristic *unemployed* had to be calculated. The variance of this characteristic is a function of the unemployment rate. The average unemployment rates from the LFS for provincial and subprovincial areas for the period from 1984 to 1992 were used. This period was chosen to reflect unemployment rates typical of those likely to be encountered during the life of the survey design. In addition, since the LFS design involves clustering, the variance can be expressed as a function of a design effect as well as the unemployment rate. [The design effect of an estimator is the ratio of its variance under the actual design to the variance it would have under a simple random sample of the same size.] The design effects used were based on estimated design effects from 1989 to 1992---smoothed averages of design effects for each ER were calculated. A more detailed explanation of how CVs were calculated is given by Mian and Laniel (1994).

Sample size reduction. Following the implementation of the new design, the size of the core sample was decreased by 6500 households. The reduction took effect in July 1995. Because of the more efficient design, the CVs of national and provincial estimates after the decrease are comparable to the CVs that existed prior to the redesign with the larger sample. The sample was reduced by the same percentage in each province. The current sample sizes for each ER and EIER are given in Table A3 in Appendix A.

Sample Selection

Stages of sampling. One of the major changes in the new LFS design is the use of only two stages of sampling in almost all areas. The first stage is an area sample. For first stage units that are selected, a list of dwellings is prepared (and maintained) in the field. A sample of dwellings is then selected at the second stage from each list.

The replacement of three stages of sampling in rural areas under the old design by a two-stage design has several benefits. In addition to being simpler, the two-stage design is more statistically efficient than the old design. Since the location of first-stage sampling units is now less constrained, the spread of the sample is improved. This can be an advantage for small area estimation; see Singh et al. (1994).

Rural areas. In the new design, within rural final strata, EAs are selected at the first stage of sampling followed by the selection of dwellings at the second stage. The EAs are selected using randomized PPS systematic (rppss) sampling, which is described in the next paragraph. PPS denotes probability proportional to size, and here, the size of a sampling unit is the number of households in the unit during the 1991 Census. Within selected EAs, a systematic sample of dwellings is chosen at random. Usually, ten dwellings are chosen in this way in each EA. In the old design, three stages of sampling were used in most rural areas, where primary sampling units (PSUs) consisting of groups of EAs were selected first. Next, EAs were selected within PSU, followed by dwellings within EA. This three-stage approach was most useful when most interviews were conducted in person since the PSU corresponded roughly to one interviewer's assignment and was convenient for travel. Since five-sixths of interviews are now conducted by telephone, a simpler design is feasible in most parts of the country, and the old PSU stage has been eliminated except in some remote areas. A comparison of design alternatives for rural areas that was done in the course of the current redesign project is given by Mantel et al. (1994).

The randomized PPS systematic method of sample selection, developed by Hartley and Rao (1962), has been

used in the LFS since the 1970s. In this method, the first stage units in a stratum are put in random order and then a PPS systematic sample of the desired size is selected. In the new design, the first stage unit is the EA, and six EAs are selected per stratum. PPS systematic sampling is described by Cochran (1977).

In the rural areas with the lowest population density, a different design was used. Geographically compact first stage units consisting of six EAs were formed and two or three such first-stage units were selected using randomized PPS systematic sampling. In the selected first stage units, a systematic sample of dwellings was selected in each of the six EAs, i.e., no sub-sampling of EAs took place. This clustered design was adopted to ensure that there would be sufficient work to occupy an interviewer in sparsely populated areas that fell in the sample.

Major urban areas, non-apartment frame. In urban areas, the first stage involves the selection of clusters. Since the Street Network File does not provide complete coverage of cities, particularly their outskirts, clusters could not be formed everywhere, and EAs or parts of EAs were used as first stage sampling units. At the second stage, a systematic sample of dwellings was selected within cluster. In the apartment frame, the apartment building plays the role of a cluster. Thus in both urban and rural areas, a two-stage design is the norm.

The selection of clusters and EAs in urban areas is done using the random group method due to Rao, Hartley and Cochran (1962); see also Cochran (1977). This method was introduced in the 1970s because it is amenable to the relatively straightforward revision of cluster selection probabilities. Such revisions may occur between major redesigns in parts of cities that have undergone rapid population growth. The method is also flexible for dealing with the changes in the LFS sample size that take place from time to time. We now present an overview of how the Rao-Hartley-Cochran (RHC) random group method is implemented in the LFS. Additional details can be found in Singh et al. (1990).

For a stratum in which the RHC method is used, the clusters are assigned at random to six groups called random groups. The number of clusters in each random group is made as equal as possible, i.e., it varies by at most one cluster. In some cases, a multiple of six groups is used. In each random group, one cluster is selected with probability proportional to size; e.g., if a cluster is twice as big as a second cluster, then the first cluster has twice the probability of being selected as the second cluster.

Within selected urban clusters in non-high income strata, a systematic sample of dwellings is selected. In Montreal, Toronto and Vancouver, six dwellings per cluster are selected. In other urban areas, eight dwellings per cluster are selected. For clusters in high income strata, a systematic sample of four dwellings is selected.

Major urban areas, apartment frame. In each apartment frame stratum, which is an open-ended list of buildings, apartment buildings are selected using PPS systematic sampling (randomized PPS systematic sampling is used in the low income strata). Within each selected apartment building, a systematic sample of five dwellings is selected.

Other urban areas. In almost all other urban areas, in the first stage, either clusters or EAs are selected using the RHC random group method, followed by the selection of dwellings. The number of dwellings selected per first stage unit varies, from three (for clusters) to ten (for EAs), because this design covers a broad range of urban and semi-urban areas.

In a few special cases, accounting for less than one percent of the sample, the first stage of sampling is to select two towns in a stratum using randomized PPS systematic sampling. Then a multiple of six (usually 12 or 18) clusters are selected in each town using randomized PPS systematic sampling. Finally, a systematic sample of three dwellings is selected in each cluster. Note that here, "town" can refer to an actual town, two small towns treated as one, or part of a larger town. This design was used in cases where it was not practical to obtain reasonable interviewer assignments with the other designs discussed above.

Remote areas. In the seven non-Maritime provinces, most of the northern part of each province is sparsely populated. As a result, the LFS uses a special design for these areas. With one exception discussed later, the sample is selected in two stages. The first stage consists of a sample of settlements, which we refer to as places, and EAs. Because of the long distances involved when interviewing in remote areas, places with fewer than 10 households or 25 persons are omitted from the design. Similarly, EAs with fewer than 25 households are omitted. Despite these omissions, the design covers about ninety percent of the remote population in each province.

A sample of EAs and places is selected using systematic PPS sampling after the units are sorted by number of households. Then a sample of dwellings is selected using systematic sampling. If a selected place or EA is too big

to be listed conveniently, it will be split into manageable clusters.

Quebec has two remote strata. One stratum follows the above design and the other follows a three stage design. The latter stratum contains eight towns which form the first stage units. The second stage units are clusters obtained from EAs by splitting large EAs and collapsing small EAs. The target cluster size is 100 households, with a tolerance of about fifty households. Two towns are selected using randomized PPS systematic sampling. Then three clusters per town are selected using the same method, and finally, nine households are selected systematically in each cluster.

Sample Rotation

Each month, a portion of the LFS sample is replaced. Rotation of sampling units occurs at each stage of the multi-stage sample design. The ultimate unit of selection, the dwelling, is replaced every six months, whereas higher-level units remain in the sample for longer periods of time. The determination of six months as the period for rotation of households is a trade-off between the cost of rotation and the increase in nonresponse that might occur if respondents were asked to remain in the survey for a longer period of time.

To ensure uniform interviewer workloads and to minimize the effect of any bias due to the number of months a dwelling has been in the survey, a rotation scheme was adopted whereby one sixth of the dwellings rotate each month. This is achieved by associating with each cluster a rotation number between one and six. This number determines the months in which the rotation of households (their birth months) take place: If the rotation number is 1, then dwellings in the cluster rotate in January and July, if 2, then in February and August, and so on.

Method of Rotation

Areas using the random group method: This cluster selection method was described near the end of the previous chapter. For the *initially* selected cluster within each random group, two numbers between one and the cluster inverse sampling ratio (ISR) are generated at random. The first determines a random start for systematic selection of dwellings within the cluster. The second determines the number of systematic samples of dwellings to be drawn from the cluster, that is, the number of six-month periods for which the cluster will remain in the sample.

Prior to each occasion for selecting a new sample of dwellings, the random start for the cluster is incremented by one, until the incremented value would exceed the cluster ISR, at which time the start reverts to 1. Cluster rotation occurs at the end of the randomly determined number of sampling occasions.

The random retention period for initially selected clusters is necessary to ensure that initial probabilities of selection of units are preserved over time. If, for example, initially selected units were retained until exhausted (that is, until all systematic samples of dwellings were used), this would eventually result in a sample with overrepresentation of larger units.

Cluster rotation is carried out by proceeding to the next cluster on the randomized list of clusters in the group. If the cluster rotation proceeds to the end of the list, the selection reverts to the first cluster on the list. As with initially selected clusters, the selection of dwellings is governed by a random start between 1 and the ISR which advances at each sampling occasion.

Areas using randomized PPS systematic sampling: Rotation of dwellings and clusters proceeds as described in the previous paragraph. There are a few urban strata with 3-stage sampling. The first stage selects PSUs within strata while the second stage selects clusters within PSUs and the last stage selects dwellings within clusters as usual. The same sample rotation applies to each stage of sampling. Urban PSUs can rotate within strata and clusters rotate within urban PSUs.

Replacements for initially selected units, and subsequent replacements, all remain in the sample until exhausted subject to the minimum life rule.The minimum life rule attempts to delay the onslaught of cluster rotation after the initial selection. This is most important in rural areas where rotation may require hiring new interviewers, but the concept was applied to all areas. The bias induced by lengthening the life of the initial selection is negated by shortening the life of the subsequent selection. As a result, the second unit stays in the sample until its regularly scheduled rotate-out month as if the minimum life adjustment had not taken place. Subsequent selections remain active in the sample for full life.

The following equation must be satisfied for unbiased selection probabilities:

$$K_1 + K_2 \leq R_{min} + 1$$

Here K_1 is the minimum number of starts for the initially selected unit, K_2 is the minimum number of starts for the subsequent replacement unit and R_{min} is the smallest ISR of all the units in the stratum.

The value of K_1 is selected based on the following rules, starting with a base value of b.

> If $b < (R_{min}+1)/2$ then K_1 is a random number in the range $[b,\ R_{min}-b+1]$.
>
> If $b \geq (R_{min}+1)/2$ then $K_1 = int(R_{min}/2)+1$.

An initial selection with a life r_1 that is less than K_1 will be extended by $K_1 - r_1$. The subsequent selection will be shortened in life by the same value of $K_1 - r_1$. It follows that $K_1 + K_2 \leq R_{min} + 1$.

Based on empirical evidence at the time of design, most strata were optimized with a basic minimum of four starts (two years in sample) whereas the apartment frame was optimized with a base of two starts. Since the apartment frame is an open-ended frame, the minimum life rule is applied to new selections as well, but in this instance a base of 4 is used. The choice is more for convenience to avoid new cluster listing for very short life spans.

Operational Steps in Sample Rotation. Rotation of the sample is automated within a system known as the Sample Design System, which identifies the units that are rotating in and out of the sample for each survey. These rotating units require manual processing by Sample Control staff, who identify the geographical area represented, and form lower stage sampling units where applicable.

Each different type of frame in the sample design uses cluster selection programs that create rotation records. These rotation records describe the sequence of advancing starts within clusters and the sequence of between cluster rotations corresponding to each initial selection within a group. Each record lasts the life of the group, up to a maximum of 40 random starts. The collection of these records constitutes the master rotation file, which is used to automatically rotate the sample. The master rotation file is changed as rotation patterns are updated (as in the open-ended apartment frame), and as old records become exhausted and are replaced by new ones.

The way in which the rotation of households is accomplished is as follows. Seven months before a particular survey date the design information on all clusters rotating for that date is identified. This includes new and existing clusters.

For the new clusters, Sample Control maps out cluster diagrams (F01s) as required. The completed F01s are sent to the Regional Offices (ROs) 20 to 23 weeks before the date of introduction in order to initiate the field listing. The interviewers' laptop computers are also sent new-listing control files to match these F01s. The dwellings captured by the interviewer are transmitted back to the central database.

The central database has already been fed with the random start and inverse sampling ratio to be applied to each cluster, new and existing, in the sample for a particular survey and date. A selection of households is accomplished by a systematic sample of the lists about six weeks prior to the interview week. Household level records are sent back to the interviewers' laptops for interviewing. These selected households remain in the sample for six months.

Every six months another set of starts for these clusters is sent to the database, thereby resulting in selection of new households. In the meantime starts for other rotations have been selected as well.

In the three-stage sample design areas, a rotation record lasts only as long as a PSU remains in the sample. Thus, as records become exhausted, it can be determined which PSUs need replacing. By referring to the design frame, replacement PSUs are determined. In the PSUs identified as rotating in, Sample Control determines their geographical location, and prepares maps of the area.

Ordinarily, the process of PSU rotation begins at least 30 weeks before the date of introduction of the sample. The advanced schedule is required since not all the clusters have been formed in these new areas. With the rotation of PSUs, the subsequent stages of selection, namely clusters, will have to be carried out, generating additional rotation records

Assigning Rotation Numbers. In assigning rotation numbers, the objective is to evenly distribute the expected sample take. The expected take is just the yield from sampling all clusters based on the design count of households used in creating the frame. This distribution is achieved simultaneously for the sample as a whole, and at the same time for the smallest possible geographical subsets of it. Adherence to these objectives implies the following.

- The workload of interviewers is stable, as roughly equal numbers of units are rotating each month.
- The sample is comprised of equal numbers of households having been in the sample for 1 vs. 2 ... vs. 6 occasions, nullifying time-in-sample effects as a cause of differences in estimates between areas, or over time.
- The sample is effectively divided into six equally representative parts, which may be used when sub-samples from the LFS frame are desired.

To achieve the aforementioned objectives, the initially selected clusters were assigned rotation numbers in such a manner as to balance the total expected take within each stratum and EIER. The assignment of rotation numbers was accomplished independently within each EIER, but a random starting factor was built in to distribute the expected take as evenly as possible at higher levels.

In most areas every stratum has six or a multiple of six selections so that each rotation group can have the same number of selections. In rural three-stage designs, anywhere from six to nine clusters are formed, leading to some stratum level collapses to form 6 rotation groups. In general the smallest units were combined to create a more or less even distribution. The apartment frame has a variable number of selections that are assigned rotation numbers randomly. The remote frame typically had less than 6 selections and were left out of the general picture - their actual yields are very uncertain in any case.

Due to the differences between the size and numbers of sampled units per stratum in apartment, urban and rural areas, the first step in balancing the sample take by rotation was to balance it as much as possible for the larger rural units at the EIER level. Rotation numbers were then assigned to the highly variable apartment units and then to the urban units to balance the take at the regional level.

Basically the assignment takes place with a sorted array of expected takes by rotation. For each stratum in the list of one EIER the takes by rotation are sorted from minimum to maximum. A running count of takes collected so far is similarly sorted in reverse order. The rotations are assigned by matching the rotation with the minimum take in the stratum to the rotation with the maximum value in the running count. Initially the rotations are given a small random take. At the end of the list the takes should not vary by more than the variation within any one stratum. Note that these are design takes, whereas actual sampling will vary considerably from this expected value in some cases.

Being open-ended, the apartment frame requires rotation assignment on a continuing basis. Every set of six selections in the systematic sample was assigned the six rotation numbers in a random order. No balancing is possible in this open-ended frame. For new additions to the list, a set of six randomly ordered rotation numbers is generated as required. The assignment of rotation numbers would entail choosing the next available rotation from this set.

Similarly in strata with three-stage designs, the cluster selection stage is not complete until the PSU rotates into the sample. The new PSU must be assigned rotation numbers at the time of introduction. A residue of design takes within the region is compiled without the rotating sample in question. Then the same technique is used to assign rotations to the new PSU.

Occasionally a cluster is assigned a rotation that does not follow the standard date of introduction. This is called an off-rotation selection. The prime candidate is the apartment frame. Being open-ended, a new cluster selection can occur in any month. At the same time the rotation assigned is random. Rather than wait for the 2-5 months to introduce the sample on rotation, it is preferable to send the cluster out as quickly as possible for interviewing, implying off-rotation.

Changes to the Labour Force Survey after the Redesign

After the introduction of the new LFS sample was completed in March 1995, there were two major changes to the survey. The first was the reduction by 11 percent of the sample beginning in July 1995. Because the new LFS design is more efficient than the previous one, the CVs of national and provincial estimates after the decrease are comparable to the CVs that existed prior to the redesign with the larger sample. The sample was reduced by the same percentage in each province.

The second major change was the redefinition of the Unemployment Insurance Regions by the Human Resources Development department. In 1996, the 61 UIRs were replaced by 53 Employment Insurance Economic Regions. Once the boundaries of the new regions were finalized, the sample size per region, and the resulting quality of LFS estimates, were studied. In five cases, the sample allocated to a region was too low to meet the requirements of the Employment Insurance program. As a result, the sample in these regions was increased in 1997, with a corresponding decrease in the sample size in regions which could now afford a

reduction. At the national level, the LFS sample size remained the same, but there were some shifts in sample among provinces. Table A0 in Appendix A presents an overview of the LFS after the above changes took place.

CHAPTER 4: Special Surveys and Supplementary Surveys

Many household surveys use the Labour Force Survey frame and sample to collect information. Surveys that do this by interviewing households that have also been selected for the LFS are referred to as *supplementary surveys.* Surveys that use the LFS frame to select a different sample of households are referred to as *special surveys.* For special surveys, the households are usually selected in clusters that are also being used for LFS interviews. Use of the LFS frame and sample in this way results in substantial cost savings for these surveys. Special and supplementary surveys are often sponsored by other government departments. Note that supplementary surveys can be divided into two types: those that use LFS households while they are still being interviewed for the LFS and those that are no longer being interviewed by the LFS, sometimes referred to as rotate-outs.

Survey	Data Collection Period
Canadian Travel Survey	January-December (monthly)
Employment Insurance Coverage	January
Survey of Household Spending	January-March
Survey of Labour and Income Dynamics	January and May
Adult Education and Training Survey	January
Residential Telephone Services Survey	February, May, August, November
Survey of Household Energy Use	February
Homeowner Repair and Renovation Survey	March
Survey of Consumer Finances	April
Cultural Capital Survey	April
National Population Health Survey	June, August, November (Feb '99)
National Longitudinal Survey of Children	November
Survey of Work Arrangements	November

Each of the six rotation groups of the LFS can be used to produce estimates. Typically, special and supplementary surveys use from one to five rotation groups for their sample, depending on the required level of reliability. Usually, the LFS birth rotation group, i.e., the one consisting of households being interviewed by the LFS for the first time, is avoided because the initial LFS interview takes longer to complete than subsequent interviews.

In some cases, only part of a rotation group's households are required. To achieve this, dwellings are dropped at random as in the LFS stabilization program. Selection can also take place within households by either random sampling or by screening for individuals with specific characteristics.

The following table lists some of the surveys using LFS rotations or the LFS frame in 1998.

Examples of Major Special and Supplementary Surveys

The Survey of Consumer Finances (SCF) is an annual household survey conducted in April. It is a supplement to the LFS based on all households in four rotation groups. Each household is sent a questionnaire by mail prior to the April LFS interview. The information from the household is then collected during the LFS interview using computer assisted interviewing. The major outputs of the SCF include income distributions before and after taxes, and mean and median incomes.These results have been used to derive income-related measures such as Low-Income Cut-Offs. Plans are underway to integrate the SCF with the Survey of Labour and Income Dynamics, discussed below.

The Survey of Household Spending(SHS) is a new annual household survey that replaces both the Family

Expenditure Survey (FAMEX) and the Food Expenditure Survey. The new survey is being introduced as part of the Program to Improve Provincial Estimates (PIPES). The overall objective of PIPES is to produce more reliable provincial estimates for use in the tax allocation formula for the Harmonized Sales Tax. The Survey of Household Spending will also continue to be used in its traditional role as a source of information for computation of the Consumer Price Index. The SHS is a special survey, i.e., it selects households in clusters containing LFS sample, but the SHS households are not interviewed by the LFS.

The new survey will be very different from FAMEX. The latter was conducted every four years, but the SHS will be conducted annually. In addition, the SHS's sample will be almost double that of FAMEX. As a result, the new survey will deplete available LFS clusters more quickly. The biggest difference between the surveys will be in the data collection methodology. The long, detailed FAMEX questionnaire will be replaced by a more streamlined, mixed-mode sequence of contacts with responding households.

Longitudinal Surveys. In the 1990s, Statistics Canada developed several new longitudinal surveys to obtain data that would fill certain information gaps about Canadians. The major new surveys, which all used the LFS for sample selection, are the Survey of Labour and Income Dynamics, the National Longitudinal Survey of Children and Youth and the National Population Health Survey. We will now describe these surveys briefly.

The Survey of Labour and Income Dynamics (SLID) was introduced to study the processes that influence the economic life of Canadians. The survey is used to investigate movements into and out of low income status, labour markets transitions and the relationship between family dynamics and economic well-being. The first panel of the Survey of Labour and Income Dynamics was introduced in 1993, followed by the second panel in 1996. The two panels overlap, and each panel is in the survey for six years. Thus the first panel will be replaced by a new one in 1999. Each panel initially consists of households that were recently interviewed by the LFS. During the life of a panel, individuals in the panel are interviewed up to twelve times, alternating between interviews about labour status in January and about income in May (persons can avoid the May interview by giving permission to Statistics Canada to used their administrative income tax data). Like other longitudinal surveys, SLID follows sampled *individuals* over time, even if they move to another province or out of the country.

Since SLID will be used to produce cross-sectional estimates as well as longitudinal ones, the sample in the two ongoing panels will be supplemented annually by a top-up sample to ensure that the total sample is representative of the population at a point in time. Each top-up sample will consist of 10,000 households that will be combined with the longitudinal sample (30,000 households) to produce cross-sectional estimates.

The National Population Health Survey (NPHS) was introduced to measure the health of Canadians over time. The NPHS was the first survey to use the new LFS design, beginning in June 1994, when it selected 25,000 households. These households were not interviewed by the LFS. To meet special demands, including sample buy-ins by some provincial governments, the NPHS supplements the sample selected from the LFS frame by samples selected using random digit dialing (RDD). These yielded an additional 800 households in the first wave of the survey. In the second wave, approximately 60,000 additional households were surveyed using RDD to produce sub-provincial data for three provinces.

To take seasonal factors into account, the NPHS sample is distributed over four quarters, with interviewing taking place in February, June, August and November.

The NPHS selects one member of each initially selected household for an in-depth interview and follows him or her over time. Interviews are conducted every two years, for a planned duration of 20 years. At each wave, for cross-sectional estimation purposes, basic health information is collected for all members of the household currently residing with the longitudinal respondent.

A more detailed description of the methodology of the NPHS is given by Tambay and Catlin (1995).

The National Longitudinal Survey of Children and Youth (NLSCY), which began in 1994, tracks a sample of children over many years to monitor their development from infancy to adulthood. It is a complex survey which began with LFS-based households to obtain a sample of children, and then obtained information from children's teachers and principals. Because only about 30 per cent of LFS households had children in the appropriate age

range, it was necessary to use more than six LFS rotation groups to attain the desired sample. In most cases, eight or nine rotation groups were used. In addition to the sample obtained directly from the LFS, the NLSCY also includes children from 2500 households selected in the first wave of the NPHS. The direct LFS sample and the NPHS sample included approximately 21,000 and 4000 children under the age of 12, respectively, for a total initial sample of 25,000 children. The sample, which will be contacted every two years, will be augmented each time with children in age groups not represented by the initial sample.

The NLSCY uses a variety of questionnaires and collection methods. The initial household interview was conducted in person using computer assisted interviewing. Ten and eleven year old children completed a self-administered questionnaire. Each child's teacher and principal were identified and asked to complete a questionnaire. For the latter, a mail-out/mail-back approach was used. In addition to the usual types of questionnaire, the NLSCY also administers tests to children, namely a mathematics test and a test of receptive vocabulary. For more information on the NLSCY, see Brodeur et al. (1995).

CHAPTER 5: Weighting and Estimation

Introduction

Estimates are obtained from the sample data using knowledge of the sample design and by employing estimation techniques from the theory of survey sampling. Each person in the sample receives a weight which we will refer to as the final weight. This weight represents the respondent's contribution to the total population and is used to derive estimates for all characteristics of interest. This weight is derived as the product of three factors: a design weight, which incorporates design information; a nonresponse adjustment, which compensates for nonresponding households; and a factor (the g-factor) that calibrates the sample to known population counts.

Once the estimates are derived, it is necessary to judge their reliability. Because the LFS is a probability sample it is possible to estimate the sampling error associated with each estimate. The sampling error can be used to make probability-based statements about survey estimates.

Occasionally, estimates are required for regions that were not planned for at the time of design of the survey or whose boundaries changed after the design of the survey. This is the case for Employment Insurance Economic Regions (EIERs). Estimates for these regions are required by Human Resources Development Canada (HRDC) to administer the Canada Employment Insurance Program. By using small area estimation techniques, improvements in quality can be obtained for such regions.

The purpose of this chapter is to describe the methodology used by the LFS to derive estimates and to provide the rationale behind the methods employed. This is followed by a description of the method of estimating the sampling error. A section is devoted to the method of deriving estimates for EIERs. Finally, changes to the estimation methodology between the present and previous sample designs, and the auxiliary information used in weighting are discussed.

The Design Weight

In any sample survey a target population is defined. The target population is the subset of the population that the characteristics of interest refer to. In any given sample, some members of the target population are selected and others are not. The selected members can be thought of as representing the non-selected members. In a probability sample, each member has a known probability of being selected. If that selection probability is one in fifty, say, then the member represents 50 persons in total. One could make 50 copies of the survey responses and by repeating this procedure for every member in the sample, create a "pseudo-population". This pseudo-population could be used for deriving the required estimates, since if the sample is representative of the population, then tabulations carried out on the pseudo-population will be very close to what would have been obtained had the true population been used. In practice, the records are not duplicated but rather are assigned a weight. Since this weight is determined by the sample design, it is referred to as the design weight. The design weight can be thought of as the number of times the record would have been replicated.

For the LFS, the following facts affect the details of the estimation procedure.

1) The survey uses a stratified, multi-stage design, with sampling conducted using probability proportional to size (PPS) selection at all stages except the final stage which uses systematic sampling.

2) Since the ultimate sampling units are households, the design weights in the LFS refer to households. As mentioned earlier, information is collected on every member of the household. Every person in the household is given the household design weight in order to eventually derive estimates referring to persons.

3) The LFS is a repeated survey. Once the survey is designed, the same design is used month after month until a new design is introduced. Historically, the survey has been re-designed every 10 years. It is expected that growth in the population will occur over the life of the design and appropriate adjustments are needed at the sampling and estimation stages.

4) At the time of the design of the survey, information from the most recent census is used. In this case, the design counts (e.g., the number of households in a city block) are from the 1991 census.

Given the sample allocation, the survey design determines an initial set of design weights. These weights could be used as long as the design and allocation remain

unchanged. However, because the penultimate units experience growth over time and the systematic sampling rate is fixed, this would lead to an ever increasing sample size (and ever increasing collection costs). It would also lead to large variations in interviewer assignment sizes both within the same assignments over time and between different assignments. To avoid this, two sampling methods are used to control the sample size. The two methods, sample stabilization and cluster sub-sampling (described below), change a household's probability of inclusion in the sample. It is necessary to adjust the initial design weights to compensate for these methods. The adjustment factors are called the stabilization weight and the cluster subweight.

Sample stabilization and cluster subsampling involve dropping households in order to address problems with sample size growth. Stabilization accommodates the slow growth over time that is the result of the increasing size of the population, which, if left unchecked, would lead to an increase in the sample size. Cluster subsampling accommodates isolated growth in relatively small areas that could present interviewers with work load problems.

The design weight for a particular household is equal to the inverse of the household's probability of inclusion in the sample. It is computed as the product of three factors. These are referred to as the basic weight, the cluster subweight and the stabilization weight.

The Basic Weight

When designing the survey, strata were formed by grouping together geographic units. Details of the stratification can be found in Chapter 2. From each stratum, the number of households to be selected is determined and fixed. For stratum **h** we will call this number $\mathbf{n_h}$. We also know the number of households in the stratum at the time of design of the survey. Denote this by $\mathbf{N_h}$. The stratum inverse sampling rate (ISR) is given as:

$$R_h = \frac{N_h}{n_h}$$

Because the LFS uses multi-stage sampling, it is necessary to determine the number of units to be selected at each stage of sampling. Consider the case of two stage sampling. The expected sample take based on design counts from each first stage unit (FSU) is fixed. This is called the *density factor* and for FSU **j** in stratum **h**, it will be denoted by $\mathbf{n^*_{hj}}$. The number of FSUs to select, $\mathbf{n_{1h}}$ is given by $\mathbf{n_h / n^*_{hj}}$. If $\mathbf{N_{hj}}$ is the number of households in FSU **j** in stratum **h** , the sampling rate for the FSU is $\mathbf{N_{hj}/n^*_{hj}}$. This is denoted by $\mathbf{R_{hj}}$. $\mathbf{R_{hj}}$ corresponds to the sampling interval used to systematically select dwellings in the final stage of sampling. It is sometimes referred to as the cluster ISR.

In some cases $\mathbf{n_{1h}}$ is fixed and $\mathbf{n^*_{hj}}$ is determined as $\mathbf{n_h/n^*_{1h}}$. In either case, the size of the stratum is set to obtain desirable sample sizes.

We can now determine a household's inclusion probability as the product of the selection probabilities at each stage. We use $\mathbf{R_{hj}}$ as the size measure for PPS sampling, for the $\mathbf{j^{th}}$ FSU in stratum **h**. The first stage inclusion probability for FSU **j** is :

$$\pi_{1hj} = \frac{n_{1h}}{\sum_{j \in h} R_{hj}} R_{hj}$$

The conditional inclusion probability of selecting household **k** given that FSU **j** is selected is, by definition,

$$\pi_{k/j} = \frac{n^*_{hj}}{N_{hj}} = \frac{1}{R_{hj}}$$

The inclusion probability of household **k** in stratum **h** then is

$$\pi_{hk} = \pi_{1hj}\, \pi_{k/j} = \frac{n_{1h}}{\sum_{j \in h} R_{hj}} R_{hj} \frac{1}{R_{hj}} = \frac{n_{1h}}{\sum_{j \in h} R_{hj}}$$

Note that

$$\sum_{j \in h} R_{hj} = \sum_{j \in h} \frac{N_{hj}}{n^*_{hj}} = \frac{n_{1h}}{n_h} \sum_{j \in h} N_{hj} = n_{1h} R_h$$

The inclusion probability equals the original stratum ISR, $\mathbf{1/R_h}$. In general, sample designs with equal basic weights within each stratum are called *self-weighting designs*. The

LFS is self-weighting within each stratum (with respect to the basic weight), and the basic weight is $\mathbf{R_h}$.

The Cluster Subweight

As described earlier, the LFS follows a multistage design. The penultimate units, or clusters, are sampled at a fixed rate determined on the basis of the 1991 census counts, to yield between 6 and 10 dwellings per cluster. In urban areas, new development often takes place and the number of dwellings in a cluster can grow substantially over time. When this occurs, given the fixed sampling rate, an interviewer's assignment size can grow substantially. This can affect the quality of the interviewer's work in addition to his/her ability to complete the assignment. When growth in a cluster exceeds 200%, the cluster may be subsampled using one of three methods. Each of these methods involves randomly dropping sampled households from the growth cluster, with the result that the household probability of inclusion is altered. Instead of continually re-computing the basic weight, it is easier to compute a weight adjustment and apply it to the original basic weight. This adjustment factor is called the *cluster subweight*. The three methods used to drop the households and determine the cluster subweight are the following.

Method I: Subclustering

When growth exceeds 300% and street patterns are defined well enough to delineate clusters, the growth cluster is divided into several clusters. A sample of the smaller clusters is taken, say $\mathbf{n_{2hj}}$ of them. The smaller clusters are sampled in a manner that will reduce the overall take. Let $\mathbf{R_{hj}}$ equal the sampling rate of the original cluster. The sizes of the new clusters, $\mathbf{N_{hji}}$, and the expected sample takes, $\mathbf{n_{hji}}$, give the sampling rates for the new clusters, $\mathbf{R_{hji}}$. We now determine the rate at which we have to sample the original cluster to obtain the total sample size obtained from the new subclusters. This is given by

$$R^{*}_{hj} = \sum_{i \in j} \frac{R_{hji}}{n_{2hj}}$$

The cluster subweight is given by

$$K = \frac{R^{*}_{hj}}{R_{hj}}$$

The original basic weights given to the households that get selected are multiplied by this factor to reflect their actual selection probability.

Method II: Self- Representing Cluster

When the characteristics of the growth dwellings are distinct from the remainder of the stratum or the size of the cluster is at least 20% of the size of a stratum, the cluster is first re-classified as a stratum. Call this new stratum **(hj)** . New clusters are formed within this new stratum and a sample is drawn. The sample from the growth cluster now represents the cluster itself, rather than the larger original stratum. If the design count of the new stratum is $\mathbf{N^{N}_{(hj)}}$ and the expected sample take is $\mathbf{n^{N}_{(hj)}}$, the stratum sampling rate is given by

$$R^{N}_{(hj)} = \frac{N^{N}_{(hj)}}{n^{N}_{(hj)}}$$

$\mathbf{R^{N}_{(hj)}}$ is the basic weight to be assigned to households selected from this new stratum. Since households selected from this new stratum are assigned the weight from the original stratum, $\mathbf{R_h}$, the appropriate factor is

$$K = \frac{R^{N}_{(hj)}}{R_h}$$

It is also necessary to apply an adjustment to all the sampled households in the remainder of the original stratum. Consider a stratum from which six clusters are selected. The six clusters are used to represent the full stratum's population. After removing the growth cluster from the stratum, the weights for households in the five remaining clusters must be adjusted so that they represent the remainder of the stratum.

Let $\mathbf{N^{R}_{h} = N_h - N^{N}_{(hj)}}$ be the design count of the remainder of the stratum and, if $\mathbf{n_{hj}}$ is the original expected take from the cluster that has been removed from the stratum, let $\mathbf{n^{R}_{h} = n_h - n_{hj}}$ be the expected sample take from the remainder of the stratum. The new inverse sampling rate for the stratum is

$$R_h^R = \frac{N_h^R}{n_h^R}$$

This leads to the cluster subweight

$$K = \frac{R_h^R}{R_h}$$

Method III: Cluster Subsampling

When a cluster is to be subsampled , and neither method I nor II applies, this, the simplest and most common case of subsampling, is used. First the cluster is sampled based on the original design sampling rates. This yields a set of sampled households. A second random selection is made from the sampled households. The households that remain after the second selection are interviewed while the remainder are dropped from the sample. If the cluster was originally sampled at a rate of $\mathbf{R}_{hj}$, and subsampling leads to a sampling rate of $\mathbf{R}^*_{hj}$ then the cluster subweight is

$$K = \frac{R^*_{hj}}{R_{hj}}$$

For example, if every second selected household is chosen to remain in the sample then the new sampling rate for the cluster is twice the old sampling rate. The above ratio would equal two. For households in this cluster, the basic weight will be multiplied by two to compensate for the discarded households. Due to outlier problems encountered by special surveys that use the LFS frame, the maximum value the cluster subweight can be is 3.

The Stabilization Weight

The final stage of sampling is conducted using systematic sampling at a fixed rate. As the sampling rate is employed consistently over time, growth in the population, and hence in the number of households, will lead to an ever increasing sample size and escalating survey costs. To control costs, sample stabilization is carried out. Sample stabilization is the random dropping of dwellings from the sample in order to maintain the sample at its desired level. By randomly dropping dwellings, a household's inclusion probability is changed. For example, suppose we define a stabilization area, **a**, in which households have a probability of inclusion of 1 in 200 at the time of design. If the stabilization area has a desired take of 300 dwellings, and sampling using the probability assigned yielded 350 dwellings, then 50 dwellings must be dropped. After dropping the 50 dwellings, the inclusion probability is no longer 1 in 200, but rather 3 in 700 (i.e., 1/200 times 300/350). As with cluster subsampling, it is simpler to adjust the basic weights where necessary rather than to continually re-compute them. The basic weight is retained as 200 but is multiplied by the factor 350/300 to yield the desired weight. The adjustment factor is called the *stabilization weight*.

It is first necessary to define stabilization areas. For the present design, a *stabilization area* is defined as all dwellings belonging to the same EIER and the same rotation group. For each stabilization area **a,** a base sample size is determined. This is the desired sample based on the sample allocation. The base sample size for area **a** is denoted $\mathbf{b}_a$. If sampling took place without stabilization, a number of dwellings would be selected. Call this number $\mathbf{n}_a$. If $\mathbf{n}_a$ exceeds $\mathbf{b}_a$ it is necessary to drop $\mathbf{n}_a - \mathbf{b}_a$ dwellings. This is done systematically at random. Once this is done we adjust the basic weight.

The LFS follows the rule that if a cluster has been subsampled using method III (see the previous section), then the cluster should be excluded from stabilization. No dwellings from that cluster can be dropped nor is the stabilization weight applied. Denote the total number of dwellings in stabilization area **a** excluded in this manner by $\mathbf{c}_a$.

There are two other cases when a household in a stabilization area does not receive the stabilization weight. On occasion, a group of households that were originally believed to be one household are encountered. These households, called *multiples*, are all included in the sample. As they did not have an opportunity to be excluded via stabilization, they do not receive a stabilization weight. Also, over the lifetime of a cluster, new dwellings are built and added to the cluster list of households. Again, since they were not eligible to be dropped, no stabilization weight is applied.

Once the dwellings have been dropped, the stabilization area is partitioned into sub-areas. A stabilization sub-area is the collection of strata within the stabilization area which have a common inverse sampling rate $\mathbf{R}_h$. The stabilization weights are calculated separately for each sub-area. In the notation we ignore this subtle point.

The stabilization weight to apply to households in area **a** is

$$s_a = \frac{n_a - c_a}{b_a - c_a} .$$

To conclude this section we repeat that the design weight, or inverse inclusion probability, **w** is given as the product of the basic weight $\mathbf{R_h}$, the stabilization weight $\mathbf{s_a}$ and the cluster subweight **K**.

Treatment of Nonresponse and Derivation of the Subweight

As with all surveys, the LFS experiences nonresponse. Nonresponse is classified into one of two types.

1. *Item nonresponse* occurs when only some information about a household is missing. This could mean some, but not all items are missing for one or more household members, or all information is missing for some but not all household members.

2. *Unit nonresponse* occurs when there is no information available for any members of the household.

Item nonresponse is treated entirely by imputation. For a particular missing item, a donor record is found among the respondents. The donor's responses to the corresponding missing information is used. Typically, a suitable donor is a person who has similar geographic and demographic characteristics and, for those items for which responses are available, similar response patterns. The details of the imputation method can be found in Lorenz (1995).

In the case of unit household nonresponse, if a nonresponding household had responded in the previous month, then the previous month's responses are "carried forward". This method is employed only if there was a response in the previous month (i.e., carried forward data is not carried forward again).

Finally, all remaining whole unit nonresponse is treated by the method of weight adjustment. The principle of weight adjustment is that the responding households can be used to represent both responding and nonresponding households. The design weight is multiplied by this nonresponse adjustment factor (defined below) and the result is called the *subweight*.

To carry out this weight adjustment, the sample is first partitioned into weight adjustment classes or *nonresponse areas*. The nonresponse areas are defined in such a way as to improve the chances that the respondents will have characteristics similar to those of the nonrespondents. In the LFS, the nonresponse area is defined as all households that belong to the same EIER, the same type of area and the same rotation group. Type of area refers to the type of frame the sample is drawn from (see Chapter 2). The classifications are as follows.

CMA Apartment design

CMA Regular design

Non-CMA Computer Assisted Districting Program (CADP) design

Urban EA design

Urban Cluster Design

Urban 3 stage design

Rural EA design

Rural 3 stage design

Remote area design.

Rotation group is included in the definition of a nonresponse area because it is known that both the magnitude and patterns (refusals, non-contacts, etc.) of nonresponse differ depending on how long a household has been in the survey. In the context of nonresponse adjustment, this is discussed in Kennedy et al. (1994) . One feature of the new design is the formation of high income strata. Because of their unique characteristics, high income strata are treated as nonresponse areas on their own. Note that the nonresponse areas do not overlap, and together they cover the entire target population.

Within each nonresponse area, a *nonresponse adjustment factor* is computed. The adjustment factor for a nonresponse area is given as the ratio of sampled households, weighted using the design weight to represent the number of households in the area, to responding households weighted to estimate the number of households in the area that would respond. If we denote by **n** the number of sampled households in nonresponse area **b**, and by **r** the number of responding households, then the nonresponse factor is

$$f_b = \frac{\sum_{k=1}^{n} \pi_k^{-1}}{\sum_{k=1}^{r} \pi_k^{-1}}$$

where π_k^{-1} is the design weight assigned to the household.

A value greater than two for the above weight is undesirable so when this occurs the nonresponse area is collapsed with another nonresponse area chosen so that when the pooled weight is computed, it will be less than two. The nonresponse area to collapse with must come from the same province, the same type of frame and the same rotation group (collapsing across EIERs if necessary).

This weighting factor is applied to all responding households in the area. The *subweight* is defined as the product of the design weight **w** and the nonresponse factor $\mathbf{f_b}$.

The Final Weight

In principle the subweight defined above could be used to produce estimates of the characteristics desired. However, from estimation theory, it is known that if auxiliary information about the target population is available, and this information is correlated with the characteristics of interest, then it can be used to produce more efficient estimates. Consider a sample that by chance consists of 50% women and 50% men. If the true distribution of males and females in the population is 51% women and 49% men, then the sample under-represents females. Many labour force characteristics are related to gender. For example, a higher proportion of men are employed. Adjusting the subweights so that the true proportion of each gender group is represented would lead to a better estimate. The adjustment factor computed to exploit auxiliary information is called the *g-factor*. The product of the subweight and the g-factor is called the *final weight*.

To obtain the g-factor, the LFS uses a form of the general regression estimator (GREG) based on the weighting methodology proposed by Lemaitre and Dufour (1987). Postcensal estimates of population projected to the current time period are used as auxiliary information. Specifically, the estimates used are population totals for 30 age/sex groups within each province, as well as population totals for Economic Regions and Census Metropolitan Areas. These population counts are produced each month by Statistics Canada's Labour and Household Surveys Analysis Division.

The LFS subweight is a household weight. The GREG estimator computes a final weight for each household, derived in such a manner that the sum of the final weights for sampled individuals in a particular age/sex group, or in a particular sub-provincial region, agree identically with the population estimates used as auxiliary information. Also, the estimates of *employed*, *unemployed* and *not in the labour force* will sum to the population totals used as auxiliary information since everyone in the sample belongs to one of these three categories. Because the weight is the same for all persons in the same household, family level estimates and person level estimates are consistent. This was not the case in the methodology employed before the Lemaitre-Dufour form of the regression estimator was adopted.

To conclude, the following are some advantages to using the final weight step:

- consistency of estimates with demographic estimates of population.
- an adjustment for coverage error.
- a common weight for all members of the same household.
- reduction in sampling error of estimates.

Algebraic Description of Weighting a Record

The following is an algebraic description of weighting. We begin by introducing notation. The LFS sample design consists of a nested hierarchy of geography.

Let p = 1, ... , 10 denote the province.

u = 1, ... , U denote the EIER u, within province p.

f = 1, ... , F denote the type of frame within EIER u.

h = 1, ... , H denote stratum h within frame f.

r = 1, ... , 6 denote the rotation group within stratum h.

j = 1, ... , J denote cluster j of rotation group r.

k = 1, ... , K denote household k in cluster j .

i = 1, ... , c_k denote individual i within household k.

With this notation a household is identified with the subscript **pufhrjk** . A subscript containing periods or missing subscripts indicates a reference to a level of accumulation. For example, **pu..r** refers to all households in province p, EIER u, and rotation group r, collecting households over the missing subscripts.

In a few cases, strata cut across EIER boundaries. For the most part this has occurred because HRDC redelineated its EIERs after the redesign of the LFS. Special estimation techniques are used to produce estimates for these regions. We note that the above geography is not quite perfectly nested. This will present no problem for the standard estimation methods discussed in this section.

At the time of design of the survey, the inverse selection probabilities are the same for all households in the same stratum. The basic weight can be denoted as

$$w_{pufh}$$

The next two weighting factors, the cluster subweight and the stabilization weight adjust the basic weight to account for various adjustments to the sample yields as described earlier in this chapter. The method of computing the cluster subweight depends on the method of subsampling employed.

Method I: Area Subsampling

In this case, the cluster is redelineated into smaller clusters. A sample of clusters is then selected and sampled to obtain some fixed total yield. If the sampling rate from the original cluster was $R_{pufh.j}$, and if the sampling rate at which the original cluster had to be sampled in order to obtain the new total sample yield is $R^*_{pufh.j}$, then the cluster subweight is

$$c_{pufh.j} = \frac{R^*_{pufh.j}}{R_{pufh.j}}$$

Method II: Self Representing Cluster

In this case, the growth cluster is removed from the stratum and forms a new stratum. The new stratum, **h'** say, is delineated into clusters and sampled. If the new strata were sampled at the same rate as the old strata, no adjustment weight would be required. However this would likely yield a very small take. If the sampling rate of the original stratum is R_{pufh}, and the sample rate of the new stratum is $R^*_{pufh'}$, then the cluster subweight to be assigned to households in the new stratum only is

$$c_{pufh'} = \frac{R^*_{pufh'}}{R_{pufh}}$$

It is also necessary to adjust the remainder of the clusters in the old stratum to compensate for losing a cluster. Recall the sampling rate of the original stratum is R_{pufh}. Denote by R^R_{pufh} the rate at which the remainder of the original stratum would be sampled to get the expected design sample take from the remaining clusters. This leads to the following factor which is applied to all households in the remainder of the stratum:

$$c_{pufh} = \frac{R^R_{pufh}}{R_{pufh}}$$

Method III: Cluster Subsampling

In this the simplest case, the selected households are subsampled and only the subsampled households interviewed. If $R_{pufh.j}$ is the original sampling rate for the cluster and $R^*_{pufh.j}$ is the cluster sampling rate required to achieve the appropriate level of subsampling, then the cluster subweight is

$$c_{pufh.j} = \frac{R^*_{pufh.j}}{R_{pufh.j}}$$

As outlined earlier, stabilization weights are computed within stabilization areas. In the present design, a stabilization area is defined as the set of all strata belonging to the same EIER. This area is then divided into common rotation groups. Within each stabilization area, a base sample size is determined. This is the number of households the area should sample based on the sample allocation. This number is denoted as $b_{pu..r}$. When sampling takes place, a realized number of households is encountered, say $n_{pu..r}$. If $n_{pu..r} > b_{pu..r}$ then the area is being over-sampled and the excess households are dropped at random, using systematic sampling. As clusters that were subsampled using Method III of cluster subsampling above are not eligible for stabilization they are excluded when computing the stabilization weight. Denote the total of these dwellings in a stabilization area as $c_{pu..r}$.

When an area is subject to stabilization the following factor is applied to households in that area:

$$s_{pu..r} = \frac{n_{pu..r} - c_{pu..r}}{b_{pu..r} - c_{pu..r}}$$

Note that some households in a stabilization area do not receive the stabilization weight. These households are defined earlier in this chapter. Essentially, they are households that were not eligible to be dropped via stabilization.

We can now compute the design weight for each household as

$$\pi^{-1}_{pufhrjk} = w_{pufh} \times c_{pufh.j} \times s_{pu..r}$$

The design weight is the inverse inclusion probability for the given household. When referring to the design weight in the following, the cumbersome subscripting will be dropped. That is,

$$\pi^{-1}_{k} = \pi^{-1}_{pufhrjk}$$

The next adjustment is the nonresponse adjustment. Nonresponse areas are defined and an adjustment weight applied to compensate for complete household nonresponse. The LFS defines nonresponse areas as all sampled households belonging to the same EIER, the same type of frame and the same rotation group. The adjustment factor is computed as the weighted ratio of sampled households to respondent households. That is,

$$f_{puf.r} = \frac{\sum_{k \in s} \pi^{-1}_{k}}{\sum_{k \in r} \pi^{-1}_{k}}$$

where summation over s indicates summation over all households in the nonresponse area and summation over r is over all responding households in the area. All households in the same nonresponse area receive the same nonresponse adjustment factor.

The subweight is given as the product of the design weight and the nonresponse adjustment:

$$a_k = f_{puf..r} \times \pi^{-1}_{k}$$

Note that all members of the same household receive the same value of the subweight.

As mentioned earlier, we could use the subweight to estimate the desired characteristics. Given a characteristic Y, *employment,* say, we are interested in the total number of persons employed in the population. This can be denoted as

$$t_y = \sum_{U} y_i$$

where summation over U indicates summation over all persons in the in-scope population (the subscript **i** above refers to persons) and y_i has a value of one if an individual **i** is employed and a value of zero otherwise.

The survey estimate based on the subweights defined above would be

$$\hat{t}_{ya} = \sum_{s} y_i a_i$$

where summation over **s** indicates summation over sampled persons only, and a_i is the subweight.It is useful to note that, in certain cases, we could re-write the above formulae as

$$t_y = \sum_{k=1}^{N} \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{N} y_k$$

and

$$\hat{t}_{ya} = \sum_{k=1}^{n} a_k \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{n} y_k \, a_k$$

where $\mathbf{c_k}$ is the number of persons in household **k** , **N** is the number of households in the population and **n** is the number of households in the sample. The $\mathbf{y_k}$ are household

totals $\sum_{i \in k} y_i$ for the characteristic of interest, such as the number of employed persons in the household. The subscript k refers to a household's total and the subscript i to an individual's value, and abusing the notation, we have used *i* instead of *ki*.

The LFS has access to post-censal population estimates that are derived independently of the sample. These are used as auxiliary information to derive a final set of weights. To exploit the auxiliary information methods such as poststratification or a regression estimator can be used. The regression approach used by the LFS is described in Lemaitre and Dufour (1987). In the following, the approach used in chapter 6 of Sarndal et al. (1992) is used.

To begin, consider the following notation.

y_i is the value of the characteristic of interest for individual i.

y_k is the household total of the characteristic of interest for household k.

Q is the number of auxiliary variables used in estimation. Each auxiliary variable will be denoted by q = 1,...,Q.

x_{qi} is the value of the q^{th} indicator variable for individual i. The indicator variable assumes a value of one if individual i belongs to the j^{th} auxiliary category and zero otherwise.

x_{qk} is the total of the values of the q^{th} indicator for all persons in household k.

x_k is a Q x 1 vector whose q^{th} entry is the corresponding household total x_{qk} .

c_k is the size of the k^{th} household.

$\hat{t}_{ya}$ is the subweight-based estimate described above.

t_{xq} is the known population total for the q^{th} auxiliary variable.

$\hat{t}_{x_q a}$ is the subweight-based estimate for the q^{th} auxiliary variable.

Thus

$$\hat{t}_{x_q a} = \sum_s x_{qi}\, a_i$$

The regression estimator used can be written as

$$\hat{t}_{yr} = \hat{t}_{ya} + \sum_{q=1}^{Q} \hat{B}_q \, (t_{x_q} - \hat{t}_{x_q a})$$

The $\mathbf{B}_q$ will be defined below. From the above formula, we can see that the regression estimator can be viewed as the subweighted estimator plus an adjustment term. If the sample-based estimate is close to the known total for $\mathbf{x}_q$, then the adjustment term will be close to zero. If they are different, the adjustment term may be large. Use of the regression estimator produces gains in efficiency if the characteristics are correlated with the auxiliary variables.

To define the $\mathbf{B}_j$, matrix notation is used.

$$\hat{B} = (\hat{B}_1,...,\hat{B}_Q)' = \Big(\sum_{k=1}^{n} \frac{x_k x_k^t a_k}{c_k} \Big)^{-1} \sum_{k=1}^{n} \frac{x_k y_k a_k}{c_k}$$

where

$$\Big(\sum_{k=1}^{n} \frac{x_k x_k^t a_k}{c_k} \Big)^{-1}$$

is a **Q×Q** matrix. This matrix is the inverted weighted sum-of-cross-products matrix in regression estimation, and

$$\sum_{k=1}^{n} \frac{x_k y_k a_k}{c_k}$$

is a **Q×1** vector.

The above estimator can be rewritten as

$$\hat{t}_{yr} = \sum_{k \in s} y_k a_k g_k$$

where

$$g_k = 1 + (t_x - \hat{t}_{xa})^t \Big(\sum_{k \in s} \frac{x_k x_k^t a_k}{c_k} \Big)^{-1} \frac{x_k}{c_k}$$

The g_k or *g-factors* are the factors applied to the subweights to obtain the final weights.

The fact that the final weights do not depend on the characteristic y means that the same weight is used for tabulating all characteristics of interest. Also note that the g-factor is at the household level and each person in a household receives the same factor. Alternatively, we could compute the g-factors as

$$g_i = 1 + (t_x - \hat{t}_{xa})^t (\sum_{i \in s} z_i z_i^t a_i)^{-1} z_i$$

where a_i is the subweight assigned to the $\mathbf{i}^{th}$ person in the sample, and the $\mathbf{z}_i$ contain, for every person, the average of the values of the indicator variables for each person in the same household. That is,

$$z_i = \frac{1}{c_k} \sum_{i}^{c_k} x_i$$

Every person **i** in household **k** receives the same value of the indicator variable **z**, namely the household average.

Variance Estimation: The Jackknife Algorithm

The variance estimator implemented in the LFS is the jackknife. In the general case, a description of the jackknife is found in Wolter (1985). Here we outline the jackknife as it is applied in the LFS. The first step in the jackknife method is to create replicate samples from the LFS data. Within each design stratum a first stage sampling unit is selected in turn. This FSU is deleted from the sample and the subweights in the remainder of the stratum are adjusted to compensate for the deletion. We then recompute the final estimates based on the replicated sample, i.e., the provincial sample including all but the deleted FSU. By repeating this procedure for every FSU in the sample, we obtain estimates for every replicate sample, as many estimates as there are FSUs. The variability among the replicate sample estimates can be used to estimate the variance of the sample estimate. For convenience, we will refer to the FSU being deleted as a replicate.

To obtain a variance using the jackknife procedure we proceed as follows.

(i) Remove all the households from a specific replicate. Replicates will be denoted by $\mathbf{a = 1, \ldots, J_h}$. That is, the $\mathbf{h}^{th}$ stratum contains $\mathbf{J_h}$ replicates each one denoted by **a**.

The total number of replicates in the sample is

$$J = \sum_{h=1}^{H} J_h$$

where **H** is the total number of strata in the sample.

(ii) In the given stratum, for all the households in the remaining $\mathbf{J_h - 1}$ replicates, an adjustment to the subweights is made. This is done to compensate for dropping the households in the deleted FSU. The adjusted weight is

$$a_k^{adj} = \frac{J_h}{(J_h - 1)} a_k$$

(iii) Using the remaining sample, with the adjusted subweights, we recompute the final weights to obtain a new estimate of the desired characteristic. The new estimate can be denoted as

$$\hat{t}_{yr(ha)}$$

The notation **(ha)** indicates that the $\mathbf{a}^{th}$ replicate from the $\mathbf{h}^{th}$ stratum was deleted in order to obtain the new estimate. Thus the above estimate is based on all but the $\mathbf{(ha)}^{th}$ replicate.

This procedure is repeated for every replicate in the sample. This leads to **J** different estimates of the desired characteristic. The formula for the variance of the estimate of a total is

$$\hat{V}(\hat{t}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{t}_{yr(ha)} - \hat{t}_{yr})^2$$

Interest often centres on the ratio of two totals. For example, the unemployment rate is the ratio of total unemployed to total labour force expressed as a percentage. In the general case a ratio $\mathbf{100(y/z)\%}$ will use the variance formula

$$\hat{V}\left(100\frac{\hat{t}_{yr}}{\hat{t}_{zr}}\right) = (100)^2\sum_{h=1}^{H}\frac{(J_h - 1)}{J_h}\sum_{a=1}^{J_h}\left(\frac{\hat{t}_{yr(ha)}}{\hat{t}_{zr(ha)}} - \frac{\hat{t}_{yr}}{\hat{t}_{zr}}\right)^2$$

Variances of the estimates of month-to-month change and for averages over a number of months require linking the jackknife estimates over time. Consider the difference estimate

$$\hat{D}_{yr} = \hat{t}^{2}_{yr} - \hat{t}^{1}_{yr}$$

and the corresponding jackknife estimates

$$\hat{D}_{yr(ha)} = \hat{t}^{2}_{yr(ha)} - \hat{t}^{1}_{yr(ha)}$$

where the superscripts refer to consecutive months. The estimate of variance is given by

$$\hat{V}(\hat{D}_{yr}) = \sum_{h=1}^{H}\frac{(J_h - 1)}{J_h}\sum_{a=1}^{J_h}(\hat{D}_{yr(ha)} - \hat{D}_{yr})^2$$

Variance of averages are obtained in a similar way. Consider the average over **n** months,

$$\hat{A}_{yr} = \sum_{i=1}^{n}\frac{\hat{t}^{i}_{yr}}{n}$$

and the jackknife estimates

$$\hat{A}_{yr(ha)} = \sum_{i=1}^{n}\frac{\hat{t}^{i}_{yr(ha)}}{n}$$

The estimate of variance is given by

$$\hat{V}(\hat{A}_{yr}) = \sum_{h=1}^{H}\frac{(J_h - 1)}{J_h}\sum_{a=1}^{J_h}(\hat{A}_{yr(ha)} - \hat{A}_{yr})^2$$

Changes from the Previous Methodology

Several changes to LFS weighting methodology were introduced with the implementation of the new sample design. Most notable were the elimination of the rural-urban factor, a change in definition of a nonresponse area, a change in the definition of the stabilization area, and the elimination of the replication of special area records. The methods used to accommodate weighting during the phase-in of the new sample are also described below.

Rural-urban Factor. In the old LFS design, some strata in non-self-representing areas consisted of both rural and urban parts. This made it possible to have an over- or under-representation of the rural or urban population. A factor was used to adjust the subweight so that the proportion of rural and urban population in any Economic Region was the same as it was at the time of the 1981 census. In the present design, the stratification is explicit and hence the sample is representative. Therefore the factor is no longer necessary.

Nonresponse Adjustment. For the majority of unit nonresponse, the LFS uses a weight adjustment. Applying such an adjustment requires making the assumption that nonrespondents can be represented by their responding counterparts in the so-called nonresponse areas. The old method of defining a nonresponse area was as a stratum in self-representing and special areas and as the rural or urban part of a primary sampling unit in the non-self-representing areas.

Nonresponse patterns for the survey have been observed over time as part of the ongoing quality monitoring program. It has long been noted that response patterns differ among different rotation groups. A household's tenure in sample tends to affect the magnitude of nonresponse. The proportions of the different types of nonrespondents (non-contact, temporarily absent and refusal) also differ with the tenure in sample. Therefore it was decided to include tenure in survey in the definition of nonresponse area. Simply adding tenure in the old definition would have led to areas that had too small a sample take for adjustment. The definition was changed to be all households in the same Employment Insurance Economic Region, in the same type of frame and with the same tenure in sample.

Stabilization Areas. As with nonresponse areas, a change in the definition of a stabilization area was implemented. Previously, such areas were defined as all strata within a province with the same basic weight. Currently, a stabilization area is defined as all households in the same Employment Insurance Economic Region and with the

same tenure in sample. The weight adjustment is then computed within the stabilization area, pooling all strata with a common sampling rate. This change reflects the added importance placed on EIERs in the sample redesign.

Special Area Replication. The old LFS design had three frames that were small in terms of population. These were the institutions frame, the remote area frame and the Quebec remote urban frame. Combined, these frames covered about two percent of the population. The cost of interviewing in these areas was substantially more than in other areas and the sample yields were typically quite small. This latter point led to small interviewer assignments covering large areas. Because of the operational difficulties of small assignments and the fact that such small populations were involved, it was decided to ignore the subprovincial regions when sampling these areas. For example not every Economic Region had a representative sample of the special area households, though each province did. To correct for this during estimation, all the special area records were replicated across Economic Regions that had population of the corresponding type. The province level weights were then proportioned to represent the subprovincial region. For example if an Economic Region contained 10 percent of the province's remote population, the sample records for remote areas were replicated into that region and their basic weight multiplied by 0.1 .

In the new design populations in institutions are not sampled from a special frame. They are no longer treated as a special case. The remaining remote area frame is sampled in the same manner as in the old design. Replication of the records is no longer required as the impact on estimates is minimal.

Weighting During the Sample Phase-In

The new LFS sample design was introduced by replacing the old sample, one rotation group per month, over six months. Whenever households selected via the old design were due to rotate out of the sample, they were replaced with households from the new design. This process began in October 1994, with the new sample being fully implemented in March 1995. Because of changes to the LFS numbering system, new subprovincial geography, and modifications to the weighting methodology, some special considerations in weighting were required.

- No stabilization of the new sample was carried out during the phase in period.
- It was decided to stop using the rural/urban factor in October 1994. As mentioned earlier, the new sample did not require this factor. Applying it through the phase-in period to the old sample only would have led to unstable weighting factors. As old sample rotated out, the imbalance in rural/urban populations for the old sample would have been exaggerated as the new sample contribution to these populations could not be accounted for.
- Nonresponse was adjusted using the old method for the old sample and the new method for the new sample.
- Special area replication continued to be carried out for the old sample only.
- Once the subweights were computed the two samples were combined for the final weighting step.

New Developments: Composite Estimation

Until now, the fact that five-sixths of the LFS sample is common between consecutive months has not been exploited to improve estimates. It is well known that in a rotating sample design the common sample can be used to produce a better estimate of change compared to simply taking the difference between the usual estimates for two consecutive months. This improved estimate of change in turn can be used to improve the estimate of level. For example, the traditional K-composite estimator is a linear combination of the usual estimate of level, say a regression estimator, and another estimate of level obtained by taking last month's estimate of level and updating it using an estimate of change based on the common sample, i.e.,

$$est'_{(t+1)} = K \times est_{(t+1)} + (1\text{-}K) \times [est'_{(t)} + change_{common}]$$

where the prime denotes a composite estimate. Although traditional composite estimators led to improved estimates, they suffered from a number of drawbacks such as consistency of estimates. Therefore, the LFS has chosen not to implement composite estimation until now.

The paper by Singh et al.(1997) describes a version of composite estimation called modified regression (MR) composite estimation. The MR estimator is similar in spirit to, but different in detail from, traditional composite estimators. In particular, it deals with all characteristics that are to be "composited" simultaneously and takes care

of the consistency issue. The MR method has the operational advantage that it fits well into the estimation framework currently used by the LFS---the characteristics of interest enter into the estimation procedure as control totals. It also has two essential properties: each sampled household will have a single weight (i.e., the weight does not depend on the characteristic of interest) and parts will add up to the corresponding total (e.g., the sum of *employed* and *unemployed* will still equal the size of the labour force, which is not the case in the traditional approach where each variable is treated separately).

For the characteristics that are controlled in the MR process, there can be substantial improvement in efficiency as measured by their variance. For example, our studies show that for employment estimates in certain industries whose regression estimates are volatile, the gain in efficiency can exceed 40 percent. For employment and unemployment estimates at the province level, the gains are more modest but still worthwhile. For example, for Ontario the gains are five and twelve percent, respectively. For estimates of month-to-month change, the gains can be much more pronounced. For example, the variance of the estimate of month-to-month change in employment in Ontario is cut in half. For change in employment in some industries, the variance is reduced by even more. One important consequence of the latter result is that certain time series which could not be seasonally adjusted effectively in the past will be adjustable when MR is implemented, i.e., MR increases the signal-to-noise ratio sufficiently to allow the seasonal adjustment procedure to detect the seasonal pattern. Based on these encouraging results, it is planned to implement MR composite estimation in the LFS beginning in the near future.

CHAPTER 6: Data Quality

LFS Quality Indicators

LFS estimates, like those produced from any other sample survey, contain sampling errors and nonsampling errors. Accordingly, if estimates from this survey are to be interpreted correctly, knowledge of their quality is required.

In a sample survey, inferences about the target population are drawn from data collected from a single portion (sample) of this population. Results are probably different from those which would be obtained if a complete census of this population were conducted under the same conditions. Errors due to the extension of conclusions based on one sample to the entire population are known as *sampling errors*. Factors which influence sampling errors include the sample size, the variability of the characteristics studied, the sample design and the method of estimation.

Nonsampling errors, as the name indicates, have nothing to do with the sampling aspect of a survey and can occur in a census (in which all units of the population participate) as well as in a sample survey. This type of error can occur at any stage in a survey (planning, design, data collection, coding, capture, editing, estimation, data analysis and dissemination) and is primarily due to human error. Nonsampling errors can also be associated with other types of error such as errors in the population estimates used by the survey, errors in sources of information and methods to produce population projections, errors in seasonal adjustments, etc.

To ensure and monitor the quality of its data, the LFS has an extensive data quality program. A whole range of quality indicators are produced on a regular basis and carefully analysed. In the presence of unusual values, those responsible for the relevant LFS operations are immediately advised, to guarantee data quality from one survey to the next. Some indicators are also monitored in a less regular manner, since their role is to assist in the identification of long-term trends or effects, for example, the consequences of certain operational or sample design changes. This long-term information about reliability of data can be used to make changes which will improve the general quality of results and help data analysts and users, internal and external, in their work. In the following, the quality indicators produced for the LFS are presented under two headings: sampling errors and nonsampling errors.

Sampling Errors

The effect of sampling error on survey estimates is a function of a number of factors. The most obvious is sample size. All other factors being constant, sampling error generally decreases as sample size increases. In addition to sample size, sampling error depends on factors such as population variability, method of estimation and sample design. For a given sample size, sampling error is linked to a range of sample design characteristics such as stratification method used, sample allocation, choice of sampling unit and selection method used at each sampling stage for a multi-stage design. In addition, for a given sample design, the method of estimation plays an important part. Finally, even if the sample size, sample design and method of estimation used were the same, estimates of different characteristics (for which data had been collected from the same sample) would have different sampling errors, since the degree of variability would vary from one characteristic to another. These errors are usually larger for characteristics which are relatively rare or distributed unevenly throughout the population than for common or homogeneously distributed characteristics. Unemployment estimates, for example, generally have a larger sampling error than employment estimates, although both are based on the same sample.

Use is usually made of the mean-square error of one or more characteristics to measure the efficiency of the sample design and of the method of estimation. *Mean-square error* is defined as the average of the squares of the deviations of the estimated value of the characteristic from its actual value in the population. In sampling theory, for finite populations, the average of estimates from all possible samples is known as the expected value of the estimate. The difference between the expected value and actual value is known as the estimation bias. The variance of the sample estimate is the average of the squares of the deviations of the estimate from its expected value. The square root of the variance is known as the standard error of the estimate.

If the method of estimating were not biased, the expected value of the estimate and the true population value would be identical, as would mean squared error and variance. Although certain methods of estimating (such as the one used for the LFS) cause a small bias, they result in smaller mean squared errors than other unbiased methods.

One of the important characteristics of a probability sample, such as the one used by the LFS, is that the variance of an estimator (and therefore the standard error) can be estimated from the sample itself. How this is done for the LFS is explained in Chapter 5. Here, we use a simplified notation for some of the quantities described there.

The *coefficient of variation* (CV) is another important measurement of quality related to sampling error. The coefficient of variation, which is obtained by calculating the ratio (expressed as a percentage) between the estimated standard error of an estimate and the estimated value, indicates the degree of reliability of the estimate. If Y is defined as an estimate of the characteristic of interest and d as the estimated standard error of this estimate, the CV is expressed as follows: (d/Y) x 100.

The estimated standard error (d) may also be used to obtain the confidence interval associated with an estimate (Y). *Confidence intervals* are used to express precision. A confidence interval is a function of the sample data which contains the actual value of a characteristic of the observed population with a given degree of confidence. If the sampling were repeated many times, it can be asserted that 95 times out of 100, the interval $Y \pm 2d$ would contain the actual value. Under the same conditions, it can be asserted that 68 times out of 100, the interval $Y \pm d$ would contain the actual value.

To highlight the links between the various measurements of precision, let us take the following example. In March 1995, the unemployment rate for the Canadian population aged 15 and over was 10.8%, and the estimate of the corresponding standard error was 0.0016. The estimate of the coefficient of variation is therefore (0.0016/0.108) = 1.48%. The 95% confidence interval, estimated from the sample, is between 10.48% and 11.12%, i.e, 0.108±0.0032. This means that with a 95% degree of confidence, it can be asserted that the unemployment rate of the target population is between 10.48% and 11.12%.

Because of the LFS's very tight monthly publication deadlines, for any given month, CVs based on the current month are not available for immediate publication. Given the stability of CVs observed in the LFS, what is provided instead is an estimate of the coefficient of variation based on the average of CVs for the previous six-month period. These estimates are updated twice a year (January-June and July-December), and appropriate adjustment factors are applied to the averages to reflect any change that has occurred (for example, a reduction in sample size).

Data collected by the LFS make it possible to produce thousands of estimates of population characteristics. There are monthly estimates, estimates of month-to-month change, estimates of level averages and estimates of the variation in annual averages, at the national, provincial and subprovincial levels. Because of space limitations in regular and special publications, it is not possible to include direct CV estimates for all published survey estimates. However, there are look-up tables which contain CV approximations for various groups of estimates. The following table presents a few representative values of coefficients of variation for the employment and unemployment characteristics at the provincial and national level based on survey data from January to July 1997.

Observed monthly coefficients of variation for 1997

Province	Employed CV (%)	Unemployed CV (%)
Newfoundland	2.2	6.1
Prince Edward Island	1.7	6.5
Nova Scotia	1.2	5.3
New Brunswick	1.2	5.5
Quebec	0.79	3.5
Ontario	0.54	3.0
Manitoba	0.91	6.5
Saskatchewan	1.1	7.4
Alberta	0.76	5.9
British Columbia	0.90	5.1
Canada	0.32	1.72

There has been an increasing focus on the quality of estimates of month-to-month change. To reflect this, the monthly LFS press release now includes standard errors of change estimates at the provincial and national levels

for *employed* and *unemployed*. These are given for the 1997 period in the table below.

Standard error of month-to-month change, *employed* and *unemployed*

Province	SE(employed) (Thousands)	SE(unemployed) (Thousands)
Newfoundland	3	2
Prince Edward Island	1	1
Nova Scotia	4	3
New Brunswick	3	2
Quebec	18	14
Ontario	20	15
Manitoba	4	3
Saskatchewan	3	2
Alberta	9	6
British Columbia	12	9
Canada	32	24

The design effect is another quality measure obtained from the sample. It is defined as the ratio between the variance of an estimate resulting from a sample survey designed in accordance with a given sample plan and the variance of the estimate which would have resulted from a simple random sample of the same size. The design effect may be used as an index of sample design deterioration over time. The LFS computes two types of design effect, each dependent on the data used to determine it. The *sample design effect* is determined from subweighted estimates, i.e., without weight adjustment to take population totals into consideration. The *overall design effect* is calculated using final weights. The sample design effect therefore reflects sampling plan efficiency only, while the overall design effect provides an overall measurement of the strategy adopted by combining all of the characteristics of the sample design (stratification, multi-stage sampling, post-stratification and estimation). The smaller the design effect, the more efficient the design with respect to sampling variance. By observing the design effect, it is thus possible to measure qualitative changes in the plan in question over time. It must be noted that sample design effects are usually larger than overall design effects based on final weights, since they do not take the gain in precision contributed by post-stratification into account.

The LFS uses the sample design effect in conjunction with other information to decide in which areas the plan should be updated. The following table presents a few values representative of sample and overall design effects for the unemployed characteristic at provincial and national levels.

Design Effects for Employed and Unemployed -- 1997

Province	Employed		Unemployed	
	Sample	Overall	Sample	Overall
Newfoundland	2.7	0.83	1.4	1.3
Prince Edward Island	2.0	0.53	1.1	1.1
Nova Scotia	2.2	0.51	1.2	1.1
New Brunswick	2.0	0.56	1.4	1.4
Quebec	2.1	0.55	1.1	1.0
Ontario	3.3	0.50	1.2	1.1
Manitoba	2.2	0.41	1.1	1.1
Saskatchewan	2.4	0.63	1.2	1.2
Alberta	4.1	0.40	1.1	1.0
British Columbia	2.1	0.50	1.2	1.1
Canada	2.8	0.51	1.2	1.1

Nonsampling Errors

Nonsampling errors can occur at any stage in a survey and are generally caused by human errors such as inattention, misunderstanding or misinterpretation. The impact on estimates may manifest itself in bias and/or variability of estimates. If the number of observations is large or if large areas are involved, the net effect of nonsampling factors on variance may be negligible. On the other hand, its effect can be large when small areas are involved or when the characteristics being studied are rare or about sensitive issues. Nonsampling bias tends to occur in one direction. It can be attributable to interviewer training or attitude, to a fault in questionnaire design or to the method of imputation used to deal with nonresponse. All of these factors can contribute to an accumulation of errors in one direction.

Nonsampling variance and/or bias may arise from a range of sources. In the following, the focus is first made on coverage, nonresponse, vacancy, response and processing. New types of indicators, which have become available since the introduction of the computer-assisted

interviewing (CAI) mode, are also presented. With these new measures, it is now possible to identify certain parameters directly related to field interviewers and to monitor the performance of the new technology that has been adopted.

Coverage Error

Coverage errors occur when sampling frame units do not appropriately represent the target population when the survey is taking place. Units may have been omitted from the sampling frame (undercoverage), units not in the target population may have been included (overcoverage), or units may have been included more than once (duplicates). However, undercoverage represents the most common form of coverage problem. Overcoverage is not a serious problem for the LFS.

Coverage errors can occur at a number of stages in the survey: during frame design, sampling unit definition, determination of selection probabilities for sampling purposes, or data collection and processing. The LFS indicator used to measure coverage error is known as the *slippage rate*. By definition, this rate is the discrepancy between LFS population estimates (without external data, i.e., based on survey subweights) and the most recent census-based population estimates. The discrepancy is expressed as a percentage of the census-based estimate. The population estimates used in determining slippage rates can also contain errors, and these errors then become a component of slippage. In the LFS, undercoverage is manifested by a positive slippage rate. To reduce the resulting bias, sample estimates are adjusted during the estimation process to population control totals from independent sources.

The omission of dwellings or individuals in the target population, i.e., the presence of undercoverage in the LFS, can introduce nonsampling errors. A dwelling is an habitable structure meeting certain criteria. Individuals living in a dwelling represent a household. An occupied dwelling may be omitted from a cluster list for a variety of reasons: omission when the list was drawn up, structure under construction during the last check, error in cluster delineation, or erroneously classified vacant. Persons may also be forgotten in a household because the respondent does not divulge their presence or because they have been allocated a usual place of residence other than in the sampled household. Students are often missed since they live elsewhere during their studies, although their usual place of residence is in the sample. Then, errors may be introduced in estimates if people not included in the survey differ from those included. For example, if the survey misses highly mobile young persons who have a higher unemployment rate than the young people included in the survey, this biases the estimates of the unemployment rate downward. Finally, as mentioned earlier, population estimates can also contribute to slippage.

Other factors contributing to slippage have been identified. For example, population growth occurs between redesigns, usually in isolated pockets in a non-uniform manner. The sampled areas may under- or over-represent the growth that has occurred, or may correctly represent the growth. Another example: the adjustment used to compensate for nonresponse (see Chapter 5) can also have an impact on slippage. If nonrespondent households have smaller household size, and if they are represented in the sample by households of larger size, the slippage rate can be affected.

Every month, slippage rates are analyzed in detail. They are produced regularly for census metropolitan areas, economic regions, at provincial and national levels, and for twelve age-sex groups in Canada (15-19, 20-24, 25-29, 30-39, 40-54, 55+). As part of the latest LFS redesign, revised series of LFS estimates beginning in 1976 were produced using 1991 census counts, adjustment for net census undercoverage, an expanded population universe (non-permanent residents are now included) and new geography. Before the 1991 Census, population estimates did not reflect census undercoverage. The slippage rate series have therefore been revised to reflect these changes, which is why slippage rates appear to be higher than was previously the case. The table below contains average slippage rates for the 1997 calendar year.

Every month, the number of temporary dockets created by Regional Offices is monitored, and explanations must be provided if there is any substantial decrease. For the selection of its sample, the LFS holds a central database which contains a list of all eligible dwellings. Six weeks before sampling, the list of dwellings to be sampled is sent to an interviewer in the field to verify and update its content. The sample is chosen from the list at Head Office. Although the list has been updated recently, it is possible when the time comes to visit these dwellings that new dwellings have to be added to the list specifically in growth areas. These dwellings are called "temporary" because a temporary docket number is allocated to them for the current month since they did not appear on the database; the situation is restored the following month.

Average Slippage Rate (%) - Canada by Age Group and Provinces - 1997

Provinces		Average
Canada	all	9.3
	ages 15-19	6.1
	ages 20-24	15.6
	ages 25-29	16.1
	ages 30-39	9.8
	ages 40-54	8.0
	ages 55 +	6.9
Newfoundland		9.8
Prince Edward Island		11.6
Nova Scotia		8.6
New Brunswick		10.4
Quebec		8.0
Ontario		9.7
Manitoba		6.1
Saskatchewan		10.7
Alberta		7.4
British Columbia		12.4

As a last indicator of coverage quality, average household size is produced occasionally for the LFS target population by province and area type: rural and urban. The series is evaluated for fluctuations or stability.

All these indicators point to potential problems with sample coverage and make it possible to react accordingly. To remedy the situation, or slow its progression, interviewer exercises may be contemplated to reinforce their knowledge of household composition rules, a bulletin can be distributed to explain slippage or the multiple dwelling concept, or a re-listing program can be established for a number of clusters deemed to be in full expansion. Slippage will always be monitored very carefully, because in terms of quality, it translates in the LFS into reduced sample size compared to what it was when the sample design was developed; it can introduce a source of possible bias and variance in the estimates. Moreover, despite the application of a method of estimation to correct for slippage, a certain persistent estimation bias can be anticipated, since the characteristics of omitted persons and dwellings can differ from those of persons included in the sample.

Nonresponse

Each month, during the survey week, interviewers have to determine which selected dwellings contain eligible persons to the survey. Dwellings are classified non-eligible for the survey month for the following reasons:

- dwelling out-of-scope, i.e., a dwelling occupied by persons not in the target population such as members of the Canadian Armed Forces;
- dwelling vacant: unoccupied, seasonal or under construction;
- dwelling non-existent: demolished, converted to business location, mobile home moved or dwelling abandoned or erroneously listed.

When a dwelling is identified as being eligible for the survey, it is not always possible to conduct the interview for the following reasons:

- nonresponse from household: no one home, temporarily absent, interview impossible (inclement weather, unusual circumstances in the household, etc.) or refusal.

Moreover, since the introduction of CAI in the fall of 1993, a new nonresponse code has appeared. This code, which previously represented questionnaires not received in time for processing because of postal problems, has retained that connotation but is attributable to technical problems. Such problems include: hard disk crash, tape drive system defect, insufficient memory allocation, excess heat, power outages, telephone troubles, etc. Most of these cases can be solved for the next month, but given very short publishing deadlines, it is often impossible to solve them during the current month. As a result, these cases are considered nonrespondents. This component of

nonresponse has become almost negligible as experience with computer-assisted interviewing has increased.

The three following approaches are used to compensate for nonrespondent units. For households nonrespondent for the current month (i.e., unit nonresponse), information provided during the previous month will be carried forward, if it exists. This procedure cannot be applied for two consecutive months, however, and it makes it possible to process approximately 30% of nonrespondents. Nonrespondent households with no information coming from the previous month are compensated for by inflating the weight of respondent households which belong to the same rotation group, employment insurance economic region and type of area (see Chapter 5) with a factor equivalent to the inverse of the response rate. The importance of bias attributable to nonresponse is usually unknown, but it is known to be closely linked to characteristic differences between respondent unit groups and nonrespondent unit groups. Indeed, this is one of the reasons why rotation group was added as an aggregation variable when compensating for nonresponse. A number of studies have shown that nonresponse behaves differently depending on duration of participation in the survey. Since the effect of this bias increases with a higher nonresponse rate, an attempt is made to keep the response rate at as high a level as possible during collection operations.

For partial (i.e., item) nonresponse, an imputation method is used. First, deterministic imputation is applied if possible, i.e., if responses to other related questions are studied and a single value is deemed possible. Should this prove unsuccessful, a "hot deck" imputation method is used. The new imputation system selects a donor at random from current month entries which have passed the editing rules. Should the iterative process of seeking a donor fail, or should the imputed record fail to meet the editing rules, given non-consistency between data collected and those imputed after a number of attempts, the default record is completely substituted. For the wages variable, which was introduced in 1997 with the new questionnaire, "warm deck" imputation is used instead. For this variable, donor selection is not restricted solely to current month data. Values transferred from previous months are also examined, since the wages question is asked when the household is participating in the survey for the first time. It is also impossible to select abnormally elevated or low values even though they could be real.

Vacant and non-existent dwellings do not contribute to bias in the sample. However, they do increase sample variance, since they reduce the number of households in the LFS sample. An error can also be introduced if dwellings are misclassified as vacants. In the LFS, a Vacancy Check Program has been set up to obtain information related to this type of error.

Since 1993, the LFS has been subject to Statistics Canada standards and guidelines for reporting nonresponse rates. Every month, weighted and unweighted nonresponse rates are forwarded to Statistic Canada's central nonresponse information database, which is mandated to compile longitudinal data for a number of regular surveys. This database requires nonresponse rates at the collection and estimation stages. Before the redesign, the LFS supplied rates for the collection stage only, since these were the same as in the estimation stage. With the new questionnaire and the new production systems introduced in 1997, it is now possible to produce different rates for the collection and estimation stages.

The following table presents average nonresponse rates for 1997, plus the minimum and maximum reached during that year. The LFS nonresponse maximum is normally reached in July and the minimum in October. Since late 1993, a number of factors have disrupted the LFS nonresponse rate series. First, the introduction of computer-assisted interviewing generated a new type of nonresponse, which was virtually non-existent previously. Urbanization of the sample design (introduced gradually from October 1994 to February 1995), i.e., a larger proportion of the sample is selected in urban areas, also had an effect, although negligible, on this series, since nonresponse rates are generally higher in urban areas than in rural areas. Finally, the new sampling design necessitated the hiring of new interviewers who tend to obtain higher nonresponse rates during their first six months with the LFS. For a historical perspective on nonresponse issues in the LFS, see the paper by Sheridan et al. (1996).

Nonresponse Rate (Unweighted), Canada and Provinces - 1997

Provinces	Average (%)	Maximum (%)	Minimum (%)
Newfoundland	4.2	5.4	3.0
Prince Edward Island	3.5	4.8	2.4
Nova Scotia	6.3	7.3	4.6
New Brunswick	4.6	5.4	3.1
Quebec	5.4	6.6	3.7
Ontario	4.8	5.7	3.7
Manitoba	3.6	5.4	2.1
Saskatchewan	3.6	4.6	2.4
Alberta	4.9	6.3	3.1
British Columbia	5.7	6.7	4.5
Canada	4.9	5.5	3.8

During the tests that preceded the implementation of the CAI method, the LFS managers were concerned about the possible effects of this change on refusals to participate in the survey, due to the presence of a notebook computer at the respondent's household at the first interview, and hence more reticence on the respondent's part. This component of nonresponse was carefully examined during the tests, and no major increase that could be directly linked to the use of computer-assisted interviews was demonstrated.

The LFS refusal rates are usually very low. Canadian monthly rates vary between 1% and 2%. Refusal rates at the provincial level are ordinarily in the same range, but can go as low as 0.5% or rise as high as 3%. Indeed, for the first time in LFS history, a computerized collection mode made it possible to use automated forms that mentioned reasons for refusing to participate in the survey such as: do not have time, do not believe in statistics, too personal, against the government. Studies can now be undertaken more readily to identify these causes, and an attempt can be made to minimize the number of refusals. The case management system specific to the computerized collection mode can also indicate the number of refusals converted into respondents by senior interviewers each month. This information was previously non-existent.

Vacancy

Dwellings correctly identified as vacant or non-existent do not introduce any bias to LFS estimates. On the other hand, estimation variance is higher, since the sample includes fewer households. LFS interviewers return to vacant dwellings every month to interview persons targeted by the survey who may have moved in since the previous survey. Non-existent dwellings are simply removed from the sample frame. Particular attention is paid to the identification of vacant dwellings, since these directly influence two other indicators. Should a dwelling be coded vacant rather than temporarily absent for example, the nonresponse rate produced for the LFS is somewhat under-estimated. And the slippage rate is overestimated, since this miscoded dwelling should have been considered when this rate was being determined. Interviewers must therefore take great pains to determine whether a dwelling is vacant, and accordingly outside the survey field, or simply occupied by a temporarily absent household and therefore in the scope of the survey.

The next table presents average vacancy rates and minimum and maximum values for 1997 at the provincial and national levels. Once again, to meet Statistics Canada standards and guidelines for reporting nonresponse, weighted and unweighted rates are produced every month and forwarded to the Statistics Canada central nonresponse information database. In general, the vacancy rate is relatively stable, posting an upward trend with further distance from the last redesign, since the sampling frame is less up-to-date. After each redesign, the vacancy rate shows a downward trend. This downward trend was even more marked after the last redesign, given the more urbanized nature of the new sample design. For this

quality indicator, certain provinces stand out because of much lower or higher posted rates.

Vacancy Rate (Unweighted), Canada and Provinces - 1997

Provinces	Average (%)	Min (%)	Max (%)
Newfoundland	15.4	14.9	16.4
Prince Edward Island	20.5	18.6	23.0
Nova Scotia	16.8	15.2	18.7
New Brunswick	14.1	13.5	15.2
Quebec	14.0	11.9	15.8
Ontario	10.8	10.0	11.3
Manitoba	17.1	16.4	17.7
Saskatchewan	14.7	12.5	15.5
Alberta	8.7	8.1	9.8
British Columbia	9.5	8.7	9.8
Canada	13.0	12.2	13.5

Response Error

This error may be attributable to questionnaire design, question formulation, respondent understanding, the manner in which the interview is conducted or the general conditions under which the survey is conducted. Response errors can occur when information is provided, received or entered in the portable computer. However, the computerized collection mode makes it possible to reduce some of these errors, since now certain editing rules are incorporated in the collection instrument and conflicts must be resolved during the actual interview. Nevertheless, the respondent may misinterpret the question, not know the answer, have forgotten or prefer to communicate a distortion of the facts for personal reasons. Moreover, interviewers may tend to explain answers or interpret them differently. Response errors, like other categories of error, can have variance and bias.

Proxy responses obtained when collecting information from a household member about another member can also lead to response errors.

In repeated surveys, where the sample is made up of a number of panels or rotation groups, the expected value of estimates may vary slightly from one rotation group to another. This leads to what is known as rotation group bias. As regards the LFS, this bias, as measured by the rotation group index, generally peaks for the one-sixth of the sample in its first interview. The rotation group index is the ratio between an estimate calculated for the sample portion participating in the survey for a certain number of times (first month, second, etc.) and the estimate calculated for the entire sample.

Brisebois and Mantel (1996) computed a modified rotation group index which adjusts for differences in the sampling error effects for the six month-in-sample groups. Their study, which is based on sample weights before adjustment for demographic controls, found several statistically significant differences among rotation groups, but their practical effect was minor. In practice, published estimates are based on weights that have been adjusted for age-sex and geographic population controls. In addition, the weights for each rotation's groups sample are adjusted to the total provincial population. Finally, nonresponse adjustment is now done separately by rotation group, unlike in the period studied by Brisebois and Mantel. As a result, the effect of rotation group differences based on these final weights is likely to be even smaller.

Processing Errors

Processing errors can occur at various stages in the survey: data capture, editing, coding, weighting or tabulation.

Using a computerized collection mode makes it possible to avoid routing errors on the questionnaire, since the application determines the next question to be asked, taking previously entered responses into account. In addition, certain editing rules are incorporated in the collection system, making it possible to detect and correct some discrepancies at the time of the interview. However, it was not possible to incorporate all the editing rules in the computerized collection instrument, since a compromise had to be reached between interview length, computer speed and computer efficiency. Head Office completes this step by applying a set of batch editing rules.

The field edit module provides two types of quality indicators: edit failure rate and edit discrepancy rate. The edit failure rate represents the percentage of forms with at least one discrepancy. A discrepancy is defined as an entry that was blanked out, modified or inserted into a blank field after performing certain edits to check for validity. The edit discrepancy rate represents the

percentage of discrepancies on a form compared to the total number of entries on the form. Edit failure rates of approximately 1% and 5% are observed for demographic and other items, respectively. For edit discrepany rates, the corresponding figures are 0.1% and 1%.

Automatic and manual coding of occupation and industry are carried out at Head Office. During the first interview, or in the presence of any change for these variables, the interviewer gathers information describing precisely the kind of business, industry or service where the person works, and information indicating clearly and accurately the kinds of work or duties. The first type of information is used to determine type of industry, while the second type serves to identify the occupation. One of the first processing operations at the Head Office consists of coding automatically the descriptive information collected for the industry and occupation variables according to the standard classification for these variables, namely SIC and SOC. Records that cannot be coded by the automatic system are coded manually by a team of LFS coders. Approximately 14,000 records are coded manually every month. To control manual coding quality, a statistical quality control plan is applied every month. The three measures used to determine the efficiency of this control process are verification rate, which is the percentage of all control-subject entries verified, average incoming quality (AIQ), which is an estimate of the percentage of records containing errors before quality control, and average outgoing quality (AOQ), which is an estimate of the percentage of entries still containing coding errors after quality control. In 1995, the average values of these three measures were 19.6% for the verification rate, 7.9% for AIQ and 4.8% for the AOQ.

To avoid errors likely to occur in estimation and tabulation, this stage is followed by a detailed examination of the result of these operations, analysis of the various diagnostics produced automatically by the system, and comparison with other data sources.

New Error Measurements

Adoption of the computerized collection mode in the fall of 1993 generated a whole range of data that had previously been difficult to access. The CAI mode features a case management system whose primary functions involve routing cases, reporting and assisting interviewers in administering the survey. All case activities are now recorded. Accordingly, during every survey month, files containing a host of information on what is happening in the field are directly produced by the case management system. For example, it is possible to obtain average time per personal or telephone interview, number of attempts to reach a respondent, best times (hour and day) for an interview, etc. It is also possible to check whether collection procedures are being followed to the letter by interviewers. For example, nonresponse codes meaning "no one home" must only be used at the end of the survey week. This new system makes it possible to determine more about the work being done in the field, and to react to questionable cases (such as, for example, interviews completed in under a minute), and accordingly, to minimize certain errors that can slip by. This new information can also be used to improve the interviewer training program and reinforce certain behaviours such as task planning or work scheduling. This database also contains the number of cases converted from refusal to respondent, which was not available previously. The case management system records all the activities undertaken in a case. A working group is now developing an operational report designed to inform Regional Offices (ROs) about any recorded situation deemed unusual during the survey month (e.g., interviews before the survey week, insufficient time between two personal interviews, night call, etc.). To be effective, this report must be produced as soon as possible after the survey week to deal with the memory effect when looking for causes in contacting interviewers.

New case management data also contain more information about editing rules followed during interviews. With the new system it is, for example, possible to determine how many editing rules have been overridden after confirmation with the respondent, how many times an editing rule has been applied and how many times an observation has failed the rules.

With the introduction of the new questionnaire and the new computer system, it has also become possible to produce an overall imputation rate, per questionnaire and per question, which was impossible before.

All these new measures will clarify what is happening in the field and help in the interpretation of results.

LFS Data Quality Reports

A number of reports are available to LFS data users and everyone involved in the survey, at Head Office in Ottawa and in the regions. Moreover, these documents are consulted regularly by members of the LFS Data Quality Committee to monitor survey quality. In general, they contain a range of quality indicators at different geographic levels and for varying periods.

Monthly LFS Operations Report. Every month, the LFS Data Quality Unit produces a report on current

month survey data quality. This report is discussed on the day before the survey data release. Its primary aim is to permit control of operations quality in the field, which is why most quality indicators are presented by Regional Office. Some series are also presented for a 26-month period to depict seasonal and monthly changes in comparison with a previous year. The report contains the following quality indicators: nonresponse rate (by RO, component, number of months in survey, area type), vacancy rate (by RO, area type), slippage rate (by province and age-sex group), sample size, number of technical problems and number of temporary dockets.

Variance Tables. The variance tables, produced every month, contain estimates, CVs, variances and design effects of the main LFS characteristics at a range of geographic levels. In addition to these indicators, the booklet also contains average household size and slippage rates at more detailed levels than those presented in the LFS Operations Report. At present, some thought is being given to opting for a modernized, exclusively electronic format.

Quality Report. The LFS Quality Report is produced twice a year, covering the January-June and July-December periods. The objective of this report is to present an in-depth examination of quality measures associated with the LFS for the six-month study period. Quality measures are also analyzed over a 30-month period to identify the trends or effects of certain operational or sample design changes. This document focuses on a number of quality indicators presented in the Operations Report, but this time at the provincial rather than regional level. In contrast to the Operations Report, the Quality Report does include text commenting on the various tables and charts. In addition to these regular indicators, there is an occasional special chapter dealing with a subject of particular interest.

Coefficient of Variation Updates. In line with Statistics Canada policy on dissemination quality standards, the LFS produces monthly variances corresponding to the estimates it produces. Six-month CV averages are updated twice a year.

Special Reports. In addition to reports produced regularly to ensure and control the quality of LFS data, certain special reports are written occasionally. An example is the report dealing with the impact on LFS data quality of the introduction of CAI (Simard and Dufour, 1995). Since this introduction was gradual, it was possible to compare the portions of the sample interviewed with computer-assisted and paper and pencil interviewing modes with respect to a number of quality indicators. The amount of significant data on the subject gave rise to a substantial special report. This was also the case for the report dealing with the impact on LFS data quality of the phase-in of the new sample (Dufour, Simard, Allard and Ray; 1996).

Quality Assurance Programs

Over the years, the LFS has equipped itself with a number of programs to ensure the quality of the data it disseminates. Some of these programs have had to be dropped because of budget restrictions. Eight programs are dealt with in the following.

Recruiting. Before applicants are hired as interviewers, their aptitudes and ability to complete survey documents are evaluated. Even before training begins, they are sent copies of documents which introduce Statistics Canada interviewing work, outline the responsibilities, techniques and skills required of an interviewer, describe Statistics Canada organization and present the portable computer as a collection instrument.

Training. The initial training period for interviewers extends over two months. It begins with a three-day classroom course where newly-hired interviewers are shown how to use their notebook computer and computer equipment, and how to complete survey forms and administrative documents. They also carry out practical exercises, simulate interviews and learn interviewing techniques.

Interviewers then receive two days of workplace training during the first survey week in which they work, and a day or two during the second week as required. During this time, a senior interviewer accompanies and observes them, indicates how to conduct interviews, and sets an example by conducting interviews him/herself. Interviewers also participate in special group training and retraining sessions, at least once a year.

Interviewers' work is evaluated as part of other programs, which will be described below. Depending on individual performance, it is determined whether they need further self-instruction courses or revision exercises to clarify certain points or remedy weaknesses.

Observation. The observation program is designed to minimize potential interviewer errors by giving senior interviewers an opportunity to observe those who report to them, to evaluate their performance and to identify problems. Each interviewer is observed at least once every 24 months. Regional Offices decide who will be

observed and when, in such a way that it is not possible to guess the order in which the observation program will take place. Apart from this program, a senior interviewer can observe an interviewer if a specific problem is suspected. The senior interviewer accompanies the interviewer for an entire day and observes how personal and telephone interviews are conducted. On the second day, the senior interviewer checks the cluster list. Later, the senior interviewer sends the results of the observation to the RO and writes periodic reports for Head Office. The senior interviewer forwards the interviewer's performance evaluation to him/her as soon as possible after the observation.

Performance Feedback. Every month, interviewer performance reports are sent to ROs. These reports deal with costs, rejection rates on editing and response rates. Senior interviewers are regularly in contact with their interviewers and bring to their attention the results of a range of performance indicators.

Vacancy Check Program. The purpose of the vacancy check program is to monitor interviewers' field work. A sample of vacant coded dwellings is selected at least once every 24 months for each interviewer. A senior interviewer returns to these dwellings to check whether the dwellings were actually vacant, and the dwellings are recoded (vacant or otherwise). Following on the results of this program, the interviewer will receive additional training when necessary. This information is also used to measure how many households coded outside the survey field contributed positively to slippage, although it is very difficult to extrapolate for the complete sample, since the selection of checked dwellings is left to RO discretion.

Validation Program. The validation program was designed to monitor interviewer performance and to provide interviewers with correctional feedback in the form of additional training depending on identified weaknesses. Interviewers are validated randomly in such a way that during the course of a year, each of them is selected twice. Approximately 2% of households are part of this program every month (except April and December when the program does not take place). The week after the survey week, senior interviewers recontact the persons who provided information during the survey week for the sample involved in this program. They ask polite questions such as confirmation of address, memory of participating in the survey, interview time, interviewer attitude, etc. The senior interviewer also takes the opportunity to thank respondents for their participation and time. This program was suspended during the introduction of the CAI collection mode, but will return in computerized form.

Re-interviewProgram. The Labour Force Survey re-interview program is designed to measure the response variability associated with data collected by the survey and to determine the causes of response errors and conditions under which these errors are likely to occur. In addition to respondent and interviewer errors, the term "response error" includes errors due to questionnaire flaws, such as awkwardly phrased questions or inaccurate response codes. This program was suspended in the early 1990s and at the moment, its future is uncertain.

Periodic analysis of data obtained from this program made it possible to discover the causes of errors and to improve all methods used on an ongoing basis, as there were impacts on a range of survey design elements such as questionnaire composition, interview techniques, interviewer training and still more. Following is an explanation of the program.

Every month, except December, a number of dwellings in the LFS sample were chosen for the re-interview program. This subsample was selected so that part of each interviewer's assignment was re-interviewed at least twice a year. RO supervisors, or senior interviewers, carried out the re-interviews at the selected dwellings by telephone during the week following the survey week, taking into account the fact that it had been two weeks since the reference week.

The re-interview sample was divided into two parts. For one of them, the senior interviewer compared responses obtained on the second interview with those provided to the interviewer the first time, and if there were differences, determined the correct response with the respondent's help. For the other part of the sample, the senior interviewer merely conducted an independent interview. Data collected from the part of the sample which was the subject of comparison were used to obtain a measurement of response bias, the hypothesis being that the interviewer and senior interviewer would not have made the same mistakes and that comparison made it possible to discover the correct answers. Data collected from the part of the sample which was not the subject of comparison were used to obtain a measurement of response variance. Comparison of responses also permitted the senior interviewer to check interviewers' work. After the re-interview, the senior interviewer discussed discrepancies in results with the interviewers, advised them and pointed out the questions with which the greatest number of errors were associated.

Cluster Yield Monitoring. LFS cluster yield is monitored on a monthly basis to detect any divergence between the number of dwellings enumerated in the field

and the number of dwellings used in developing the sample design. The sample design uses a number derived from counts available from previous census data. Accordingly, any major difference, i.e. any 50% (positive or negative) discrepancy between enumeration and derivation, is the subject of an investigation. First, any cluster with an unexpected yield is brought to the attention of the unit responsible for sample control in Ottawa, which checks cluster boundaries and the expected number of dwellings. If the discrepancy cannot be explained at Head Office, the cluster is forwarded to the Regional Office concerned for detailed analysis. All causes explaining discrepancies are stored for future reference.

This monitoring plays an important role, since if the sample size necessitates changes, it is essential to determine which areas are being under- or over-sampled. Moreover, recorded discrepancies can reveal survey problems that could mar the quality of LFS data.

LFS Committees

The LFS has a number of coordinating groups to ensure continued good functioning of the survey. Certain committees play a permanent role at the LFS, while others are active only during the redesign. During the last redesign, there were two main committees : the Redesign Steering Committee and the Redesign Operations Committee. The first of these two Committees, chaired by the director general of the Labour and Household Surveys Branch, was mandated to oversee the overall redesign project, i.e., to monitor activities related to the introduction of CAI, the sample redesign, the new questionnaire, new processing systems and new outputs. The Redesign Operations Committee, chaired by the director of the Household Surveys Division, was in charge of managing and controlling work on individual projects.

In the following sections, three permanent committees are discussed. Their mission is to deal on a regular basis with permanent operations and evaluation of the survey.

Operations Committee. This committee's mandate is to review events occurring during survey months and circumstances surrounding the carrying out of the survey, to see that operations run smoothly, to examine proposed changes and to recommend their adoption, with a view to ensuring that the survey continues to reach its objectives. The committee, which meets every two weeks, is chaired by a member of the Household Surveys Division.

Population Estimation Steering Committee. This committee's mandate is to monitor postcensal population estimates required by the LFS. This committee also evaluates data sources and methods used to obtain the estimates at different geographic levels, and initiates a number of research projects on the topic. A member of the Labour and Household Surveys Analysis Division chairs this committee.

Data Quality Committee. When this committee was officially struck in the spring of 1972, its function was to disseminate the quality of data from the LFS and its supplements. Since that time, the committee's mandate has expanded somewhat. Its current mandate is to examine and evaluate the quality of LFS data on a monthly basis, to propose and review research and development projects designed to fine-tune methods that could influence data quality, and finally to oversee research and development in this field. The committee is chaired by a member of the Household Survey Methods Division.

To ensure that LFS data are of the best possible quality, the Data Quality Committee regularly reviews the various quality indicators mentioned earlier. It meets every month to examine and evaluate the quality of monthly data and to make suggestions and recommendations on any aspect affecting the quality of data that ought to be improved. A number of documents are made available to committee members to help them in their task. By closely following the evolution of these indicators, the committee can intervene immediately with those responsible for the LFS operations concerned to control monthly data quality. The committee also discusses new facts likely to influence the quality of data that have just been collected or are to be collected in the future, particularly changes to collection methods or to the questionnaire, unusual problems in the field, ongoing testing of procedures and methods, etc.

References

Brisebois, F. and Mantel, H. (1996). Month-in-sample effects for the Canadian Labour Force Survey. SSC Annual Meeting, June 1996, *Proceedings of the Survey Methods Section.*

Brodeur, M., Montigny, G. and Bérard, H.(1995). Challenges in developing the National Longitudinal Survey of Children. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition, John Wiley and Sons, New York.

Chen, E.J., Gambino, J., Laniel, N. and Lindeyer, J. (1994). Design and estimation issues for income in the redesign of the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Dufour, J., Simard, M., Allard, B. and Ray, G. (1996). Redesign of the Labour Force Survey Sample: impact on data quality. Statistics Canada, internal document.

Drew, J.D., Bélanger, Y. and Foy, P. (1985). Stratification of the Canadian Labour Force Survey. *Survey Methodology*, 11, 95-110.

Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Kennedy B., Drew J. D., and Lorenz P. (1994). The Impact of Nonresponse Adjustment on Rotation Group Bias in the Canadian Labour Force Survey. Presented at the 5th International Workshop on Household Survey Nonresponse. Ottawa, Canada.

Lemaître, G.E. and Dufour J. (1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology,* 13, 199-297.

Lorenz, P. (1995). Labour Force Survey - Head Office Hot deck Imputation System Specifications - Version 3. Statistics Canada, internal document.

Mantel, H., Laniel, N., Duval, M.-C. and Marion, J. (1994). Cost modelling of alternative sample designs for rural areas in the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Mian, I.U.H. and Laniel, J. (1994). Sample allocation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 24, 482-491.

Sarndal, C.E., Swensson, B and Wretman J., (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Sheridan, M., Drew, D. and Allard, B. (1996). Response rate and the Canadian Labour Force Survey: Luck or Good Planning? Proceedings of Statistics Canada Symposium 96 on Nonsampling Errors, 67-75.

Simard, M. and Dufour, J. (1995). Impact of the introduction of Computer-Assisted Interviewing as the new Labour Force Survey data collection method. Statistics Canada, internal document..

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey, 1984-1990*. Statistics Canada. Catalogue Number 71-526.

Singh, M.P., Gambino, J. and Mantel, H. (1994). Issues and strategies for small area data (with discussion). *Survey Methodology,* 20, 3-22.

Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Statistics Canada (1998). *Guide to the Labour Force Survey.* Available on the internet at www.statcan.ca/english/concepts/labour/index.htm

Sunter, D., Kinack, M., Akyeampong, E. and Charette, D. (1995). Redesigning the Canadian Labour Force Survey Questionnaire: Development and Testing. Statistics Canada internal document.

Tambay, J.-L. And Catlin, G. (1995). Sample Design of the National Population Health Survey. *Health Reports*, 7, 29-38.

Wolter K. M., (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Appendix A

Table A0-1: 1998 LFS Sample by Province and other Breakdowns*

	Strata			Households				Source	
Province	Urban	Rural	Apartment	Urban	Rural	Apartment	Total	Core	EI
Newfoundland	35	13	0	1151	836	0	1987	1987	0
Prince Edward Island	17	12	0	603	818	0	1421	1421	0
Nova Scotia	46	27	2	1753	1714	90	3557	2611	946
New Brunswick	45	20	0	1800	1261	0	3061	2604	457
Quebec	209	38	12	7693	2383	283	10358	5458	4900
Ontario	298	40	34	11717	2158	1634	15508	7179	8329
Manitoba	59	14	6	2481	1152	273	3906	3255	651
Saskatchewan	56	21	1	2471	1537	63	4072	3408	664
Alberta	75	13	6	2903	914	174	3991	3991	0
British Columbia	114	12	7	3882	854	234	4969	4113	856
Canada	954	210	68	36454	13627	2751	52830	36027	16803

Table A0-2: 1998 LFS Sample by EIER and other Breakdowns*

	Strata			Households				Source	
EIER	Urban	Rural	Apartment	Urban	Rural	Apartment	Total	Core	EI
020	15	0	0	465	0	0	465	465	0
021	20	13	0	686	836	0	1522	1522	0
123	17	12	0	603	818	0	1421	1421	0
224	8	2	0	294	145	0	439	397	42
225	7	6	0	287	342	0	629	555	74
226	22	0	2	716	0	90	806	806	0
227	4	7	0	235	466	0	701	490	211
228	5	12	0	221	761	0	982	363	619
329	32	4	0	1270	225	0	1495	1090	405
330	13	16	0	530	1036	0	1566	1514	52
433	8	9	0	323	541	0	864	455	409
434	17	0	2	730	0	45	775	510	265
435	15	0	0	566	0	0	566	0	566
436	3	3	0	238	368	0	606	224	382
437	13	0	0	722	0	0	722	153	569
438	13	5	0	590	325	0	915	401	514
439	80	0	9	1971	0	202	2173	2131	42
440	20	15	0	667	670	0	1336	732	604
441	6	3	0	305	300	0	605	295	310
442	9	3	0	507	179	0	686	316	370
443	12	0	1	534	0	36	570	169	401
444	13	0	0	540	0	0	540	72	468

* Breakdowns by Core and EI are approximate.

. . . continues . . .

1998 LFS Sample by EIER and other Breakdowns* (continued)

	Strata			Households				Source	
EIER	Urban	Rural	Apartment	Urban	Rural	Apartment	Total	Core	EI
546	18	0	5	684	0	191	875	571	304
547	16	9	0	553	527	0	1080	387	693
548	29	14	0	901	561	0	1462	657	805
549	18	0	2	780	0	94	874	132	742
550	66	0	16	1971	0	827	2798	2798	0
551	14	0	4	604	0	133	737	448	289
552	17	0	1	760	0	56	816	258	558
553	15	0	2	641	0	157	798	282	516
554	7	4	0	290	185	0	475	136	339
555	11	0	2	558	0	81	638	167	471
556	17	0	2	720	0	95	815	238	577
557	5	2	0	353	227	0	580	262	318
558	20	8	0	843	529	0	1372	439	933
559	17	0	0	760	0	0	760	73	687
560	15	0	0	712	0	0	712	52	660
561	13	3	0	587	129	0	716	279	437
664	45	0	6	1622	0	273	1895	1895	0
665	10	11	0	522	742	0	1264	1139	125
666	4	3	0	337	410	0	747	221	526
767	18	0	0	767	0	0	767	562	205
768	16	0	1	709	0	63	773	623	150
769	14	17	0	581	1023	0	1604	1604	0
770	8	4	0	414	514	0	928	619	309
871	25	0	3	861	0	85	946	946	0
872	29	0	3	970	0	89	1059	1059	0
873	21	13	0	1072	914	0	1986	1986	0
975	12	4	0	530	218	0	748	748	0
976	60	0	6	1573	0	196	1769	1769	0
977	13	0	1	612	0	38	649	351	298
978	21	4	0	628	356	0	984	627	357
979	8	4	0	539	280	0	819	618	201
Canada	954	210	68	36454	13627	2751	52830	36027	16803

* Breakdowns by Core and EI are approximate.

Table A1. Stratification Results for High Income Strata in the LFS Redesign

CMA	Dwellings	Strata	Clusters	Median Income	Average Income
Montreal	15,237	3	83	$121,881	$132,818
Ottawa	6,558	2	39	$111,729	$116,973
Toronto	35,433	4	185	$144,387	$156,477
Hamilton	6,584	1	34	$101,875	$107,130
London	4,036	1	21	$108,009	$108,604
Winnipeg	7,543	2	42	$96,763	$100,264
Calgary	7,501	1	41	$123,066	$131,543
Edmonton	5,835	1	28	$111,334	$118,600
Vancouver	16,483	3	89	$119,777	$122,739
Total	105,210	18	562	$122,765	$132,217

Table A2. Provincial sample allocations and CVs under old design, redesign and current (post-reduction)

Province	Old Sample				Redesigned Sample				Current Sample			
	Core		Total		Core		Total		Core		Total	
	Size	CV	Size	CV	Size	CV	Size	CV	Size	CV	Size	CV
Newfoundland	2,240	5.4	2,582	5.0	2,240	5.1	2,240	5.1	1,884	5.8	1,884	5.8
Prince Edward Island	1,421	6.6	1,421	6.6	1,421	6.1	1,421	6.1	1,421	6.1	1,421	6.1
Nova Scotia	3,101	5.2	4,002	4.8	3,102	4.9	4,050	4.5	2,609	5.5	3,557	5.1
New Brunswick	3,095	5.3	3,441	5.0	3,096	5.2	3,480	5.0	2,604	5.7	2,988	5.3
Quebec	6,474	4.0	11,356	3.4	6,436	3.7	11,590	3.1	5,413	4.2	10,567	3.5
Ontario	8,517	3.8	17,388	3.2	8,473	3.4	17,206	2.8	7,125	3.7	15,858	2.9
Manitoba	3,276	6.3	3,897	6.1	3,869	5.0	4,428	4.8	3,254	5.6	3,813	5.3
Saskatchewan	4,527	5.0	4,563	5.0	4,053	5.0	4,107	5.0	3,409	5.6	3,463	5.6
Alberta	5,205	4.6	5,225	4.6	4,745	4.4	4,745	4.4	3,991	4.9	3,991	4.9
British Columbia	4,454	5.7	4,975	5.1	4,875	4.6	5,583	4.4	4,100	5.2	4,808	4.8
Canada	42,310	1.96	58,850	1.67	42,310	1.73	58,850	1.50	35,810	1.96	52,350	1.62

Table A3.1. Sample allocation by Economic Region, July 1995.

Province	ER	Sample	Province	ER	Sample	Province	ER	Sample
Newfoundland	10	680	Quebec	450	553	Manitoba	660	273
	20	200	(cont.)	455	595	(cont.)	670	1790
	30	473		460	685	Saskatchewan	710	957
	40	531		465	471		720	488
Prince Edward Island	110	1421		470	1186		730	878
Nova Scotia	210	397		475	732		740	430
	220	630		480/490	404		750/760	709
	230	682	Ontario	510	1455	Alberta	810	348
	240	1002		515	926		820	300
	250	846		520	455		830	1027
New Brunswick	310	683		530	3695		840	300
	320	576		540	2193		850	300
	330	837		550	1988		860	1114
	340	519		560	1102		870	300
	350	374		570	905		880	300
Quebec	410	432		580	514	British Columbia	910	1042
	415	432		590	1740		920	2423
	420	914		595	885		930	549
	425	462	Manitoba	610	320		940	200
	430	910		620	200		950	200
	435	1387		630	374		970	200
	440	1139		640	200		980	200
	445	265		650/680	657		Total = 52350	

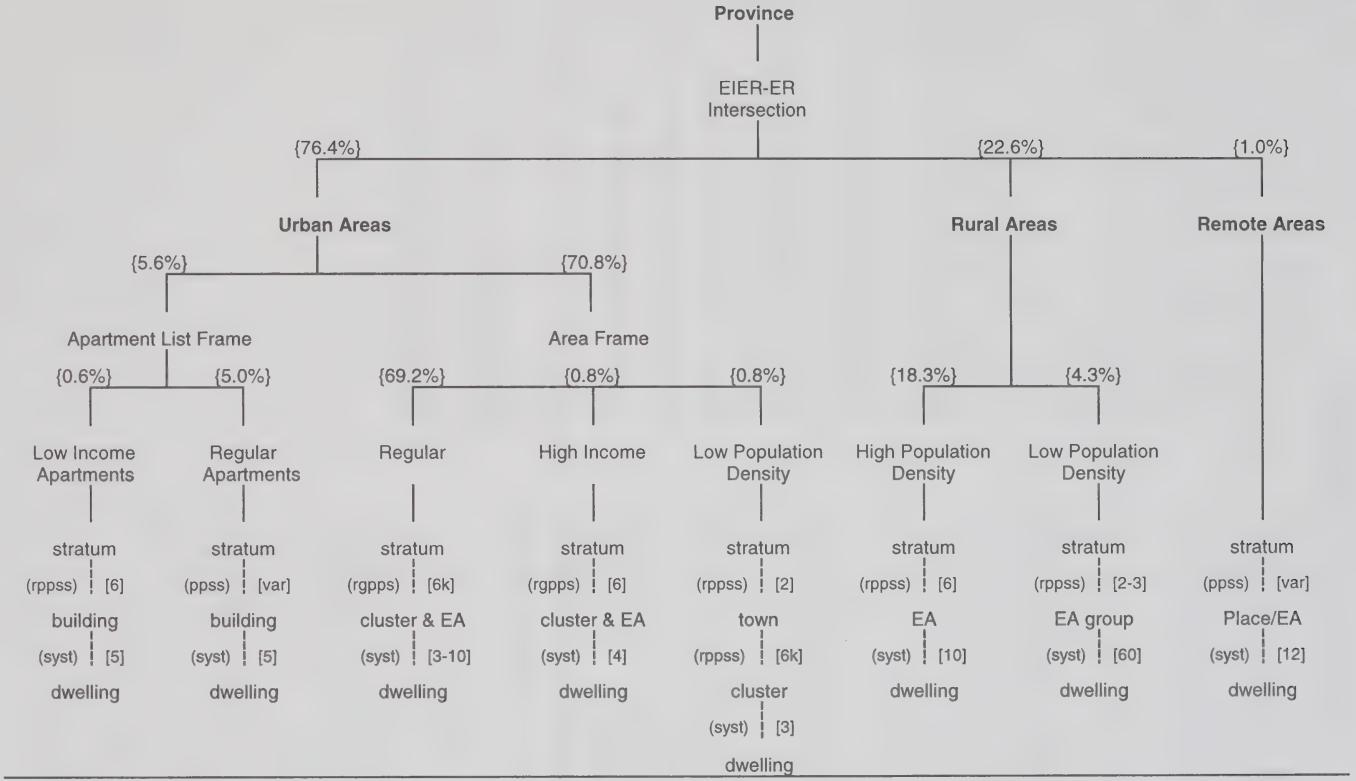
Table A3.2. Sample size allocation by Employment Insurance Economic Region, July 1995.

Province	EIER	Sample	Province	EIER	Sample	Province	EIER	Sample
Newfoundland	1	367	Quebec (cont.)	22	717	Ontario (cont.)	42	760
	2	868		23	650		43	712
	3	649		24	571		44	786
Prince Edward Island	4	1421		25	598	Manitoba	45	1920
Nova Scotia	5	439		26	538		46	1056
	6	629	Ontario	27	847		47	837
	7	805		28	858	Saskatchewan	48	616
	8	702		29	685		49	623
	9	982		30	873		50	1318
New Brunswick	10	641		31	2777		51	905
	11	592		32	738	Alberta	52	945
	12	1222		33	752		53	1057
	13	534		34	798		54	876
Quebec	14	897		35	748		55	1111
	15	778		36	539	British Columbia	56	748
	16	694		37	815		57	594
	17	564		38	815		58	1757
	18	680		39	693		59	500
	19	723		40	849		60	610
	20	1023		41	813		61	601
	21	2134					Total = 52350	

Appendix B: Abbreviations

CA	census agglomeration
CADP	computer assisted districting program
CD	census division
CMA	census metropolitan area
CSD	census subdivision
CT	census tract
CV	coefficient of variation
EA	(census) enumeration area
EIER	Employment Insurance Economic Region
ER	Economic Region
FSU	first stage unit
HRDC	Human Resources Development Canada
ISR	inverse sampling ratio
LFS	Labour Force Survey
PPS	probability proportional to size
PSU	primary sampling unit
RHC	Rao-Hartley-Cochran (random group method)
RO	Regional Office
RPPSS	randomized PPS systematic
SNF	Street Network File
UIR	Unemployment Insurance Region
VR	visitation record

Appendix C: Labour Force Survey Sample Design - 1995+



= level of stratification

= stage of sampling

EIER - Employment Insurance Economic Region
ER - Economic Region
{%} - percentage of total sample

EA - Census Enumeration Area
cluster - set of blockfaces
( ) - selection method
[ ] - number of units selected
(6k = multiple of six,
var = variable number)

pps - probability proportional to size
ppss - pps systematic
rppss - randomized ppss
rgpps - random group pps
syst - systematic

Appendix D: Urban Area Designs

Notation: E denotes EA design, V denotes VR design. The terms EA, VR and CADP are explained in the document.

CMAs: All CMAs have CADP clusters. Most also have some EA-based clusters; -E denotes absence of EA-based clusters. A few CMAs have some VR-based clusters. CMAs with apartment strata and high income strata are noted in the body of this document.

St. John's, Nfld; Halifax, NS; St. John, NB;
Quebec: Chicoutimi, Montreal, Quebec, Sherbrooke, Trois Rivières (V)
Ontario: Hamilton (-E), Kitchener (-E), London (-E), Oshawa (-E), Ottawa-Hull, St. Catharines (-E), Sudbury (V), Thunder Bay, Toronto, Windsor
Winnipeg, MB; Regina, SK; Saskatoon, SK; Calgary, AB; Edmonton, AB; Vancouver, BC (-E); Victoria, BC (-E)

CAs containing CADP clusters

New Brunswick: Fredericton, Moncton

Ontario: Belleville (some E), Brantford, Guelph, Kingston (some E), North Bay, Peterborough (some E), Sarnia-Clearwater, Sault Ste. Marie, Stratford, Woodstock

Alberta: Lethbridge, Red Deer

British Columbia: Kamloops, Kelowna (some E), Matsqui (some E), Prince George

CAs *not* containing CADP clusters

Newfoundland: Corner Brook (E), Grand Falls-Windsor (E), Gander (V), Labrador City (V)

Prince Edward Island: Charlottetown (V), Summerside (V)

Nova Scotia: New Glasgow (E), Sydney-Sydney Mines (E), Truro (E), Kentville (V)

New Brunswick: Bathurst (E), Campbellton (E), Edmunston (E)

Quebec -- EA design: Alma, Baie Comeau, Cowansville, Drummondville, Granby, Joliette, La Tuque, Magog, Matane, Rimouski, Rivière du Loup, Rouyn-Noranda, Saint Georges, Saint Hyacinthe, Saint Jean sur Richelieu, Salaberry de Valleyfield, Sept Iles, Shawinigan, Sorel, Thetford Mines, Val d'Or, Victoriaville
Quebec -- VR design: Lachute

Ontario -- EA design: Barrie, Brockville, Chatham, Cobourg-Port Hope, Collingwood, Cornwall, Elliot Lake, Haileybury, Kenora, Leamington, Lindsay, Midland, Orillia, Owen Sound, Simcoe, Timmins
Ontario -- VR design: Kirkland Lake, Pembroke, Tillsonburg

Manitoba: Brandon (E), Portage La Prairie (V), Thompson (V)

Saskatchewan -- EA design: Moose Jaw, North Battleford, Prince Albert, Swift Current, Yorkton
Saskatchewan -- VR design: Battleford, Estevan, Lloydminster, Weyburn

Alberta: Fort McMurray (E), Grand Centre (E), Grande Prairie (E), Lloydminster (E), Medicine Hat (E)

British Columbia -- EA design: Campbell River, Chilliwack, Courtenay, Cranbrook, Duncan-Chemainus, Nanaimo, Penticton, Port Alberni, Powell River, Prince Rupert, Terrace, Vernon
British Columbia -- VR design: Dawson Creek, Fort St. John, Kitimat

Other Urban Strata

Atlantic Provinces: VR design unless noted otherwise

Newfoundland: Carbonear, Channel-Port Aux Basques, Deer Lake (E), Grand Bank-Fortune, Happy Valley-Goose Bay, Marystown-St. Lawrence, Stephenville-Stephenville Crossing, Wabana-Bell Island

Nova Scotia: Amherst, Antigonish, Bridgetown-Middleton, Bridgewater (E), Digby, Liverpool, Shelburne, Windsor-Hantsport, Yarmouth

New Brunswick: Chatham-Newcastle (E), Dalhousie, Grand Falls, Oromocto, Shediac-Sackville, St. Stephen, Sussex-Sussex Corners, Woodstock

Quebec: Amos (E), Chibougamau (V), Malartic (E)

Ontario: Arnprior (V), Fergus (V), Petawawa (V), Renfrew (V), Sturgeon Falls (V)

Manitoba -- VR design: Dauphin, Flin Flon-the Pas, Swan River

Sasketchewan -- VR design: Humboldt, Melfort, Melville, Nipawin

British Columbia -- EA design: Hope, Parksville-Qualicum Beach, Trail-Rossland

Other Urban Areas: All other urban areas not included above are not strata. They are either PSUs, parts of a larger urban stratum, or are incorporated into a rural stratum.

Annexe D: Plan pour les secteurs urbains

Notation: S dénote un plan SD, F dénote un plan FV. Les termes SD, FV et PPCAO sont expliqués dans ce document.

RMR: Tous les RMR sont formés de grappes PPCAO. La majorité ont aussi des gappes basées sur des SD; E dénote l'absence de grappes basées sur des SD. Quelques RMR ont certaines grappes basées sur les FV. Les RMR avec des strates d'appartements et des strates de haut revenu sont mentionnés dans la partie principale du document.

St. John's, T.-N.; Halifax, N.-É.; St. John, N.-B.;
Québec: Chicoutimi, Montréal, Québec, Sherbrooke, Trois-Rivières (V)
Ontario: Hamilton (S), Kitchener (S), London (S), Oshawa (S), Ottawa-Hull, St. Catharines (S), Sudbury (V), Thunder Bay, Toronto, Windsor
Winnipeg, MB; Regina, SK; Saskatoon, SK; Calgary, AB; Edmonton, AB; Vancouver, BC (S); Victoria, BC (S)

AR contenant des grappes PPCAO

Nouveau-Brunswick: Fredericton, Moncton

Ontario: Belleville (quelques E), Brantford, Guelph, Kingston (quelques E), North Bay, Peterborough (quelques E), Sarnia-Clearwater, Sault Ste-Marie, Stratford, Woodstock

Alberta: Lethbridge, Red Deer

Colombie-Britannique: Kamloops, Kelowna (quelques E), Matsqui (quelques E), Prince George

AR ne contenant pas de grappes PPCAO

Terre-Neuve: Corner Brook (S), Grand Falls-Windsor (S), Gander (F), Labrador City (F)

Île-du-Prince-Édouard: Charlottetown (F), Summerside (F)

Nouvelle-Écosse: New Glasgow (S), Sydney-Sydney Mines (S), Truro (S), Kentville (F)

Nouveau-Brunswick: Bathurst (S), Campbellton (S), Edmunston (S)

Québec -- plan SD: Alma, Baie Comeau, Cowansville, Drummondville, Granby, Joliette, La Tuque, Magog, Matane, Rimouski, Rivière du Loup, Rouyn-Noranda, Saint-Georges, Saint-Hyacinthe, Saint-Jean-sur-le-Richelieu, Salaberry de Valleyfield, Sept-Îles, Shawinigan, Sorel, Thetford Mines, Val d'Or, Victoriaville

Québec -- plan FV: Lachute

Ontario -- plan SD: Barrie, Brockville, Chatham, Cobourg-Port Hope, Collingwood, Cornwall, Elliot Lake, Haileybury, Kenora, Leamington, Lindsay, Midland, Orillia, Owen Sound, Simcoe, Timmins

Ontario -- plan FV: Kirkland Lake, Pembroke, Tillsonburg

Manitoba: Brandon (S), Portage La Prairie (F), Thompson (F)

Saskatchewan -- plan SD: Moose Jaw, North Battleford, Prince Albert, Swift Current, Yorkton

Saskatchewan -- plan FV: Battleford, Estevan, Lloydminster, Weyburn

Alberta: Fort McMurray (S), Grand Centre (S), Grande Prairie (S), Lloydminster (S), Medicine Hat (S)

Colombie-Britannique -- plan SD: Campbell River, Chilliwack, Courtenay, Cranbrook, Duncan-Chemainus, Nanaimo, Penticton, Port Alberni, Powell River, Prince Rupert, Terrace, Vernon

Colombie-Britannique -- plan FV: Dawson Creek, Fort St. John, Kitimat

Autres strates urbaines

Provinces de l'Atlantique: plan FV à moins que mentionné autrement.

Terre-Neuve: Carbonear, Channel-Port Aux Basques, Deer Lake (S), Grand Bank-Fortune, Happy Valley-Goose Bay, Marystown-St. Lawrence, Stephenville-Stephenville Crossing, Wabana-Bell Island

Nouvelle-Écosse: Amherst, Antigonish, Bridgetown-Middleton, Bridgewater (S), Digby, Liverpool, Shelburne, Windsor-Hantsport, Yarmouth

Nouveau-Brunswick: Chatham-Newcastle (S), Dalhousie, Grand Falls, Oromocto, Shediac-Sackville, St. Stephen, Sussex-Sussex Corners, Woodstock

Québec: Amos (S), Chibougamau (F), Malartic (S)

Ontario: Arnprior (F), Fergus (F), Petawawa (F), Renfrew (F), Sturgeon Falls (F)

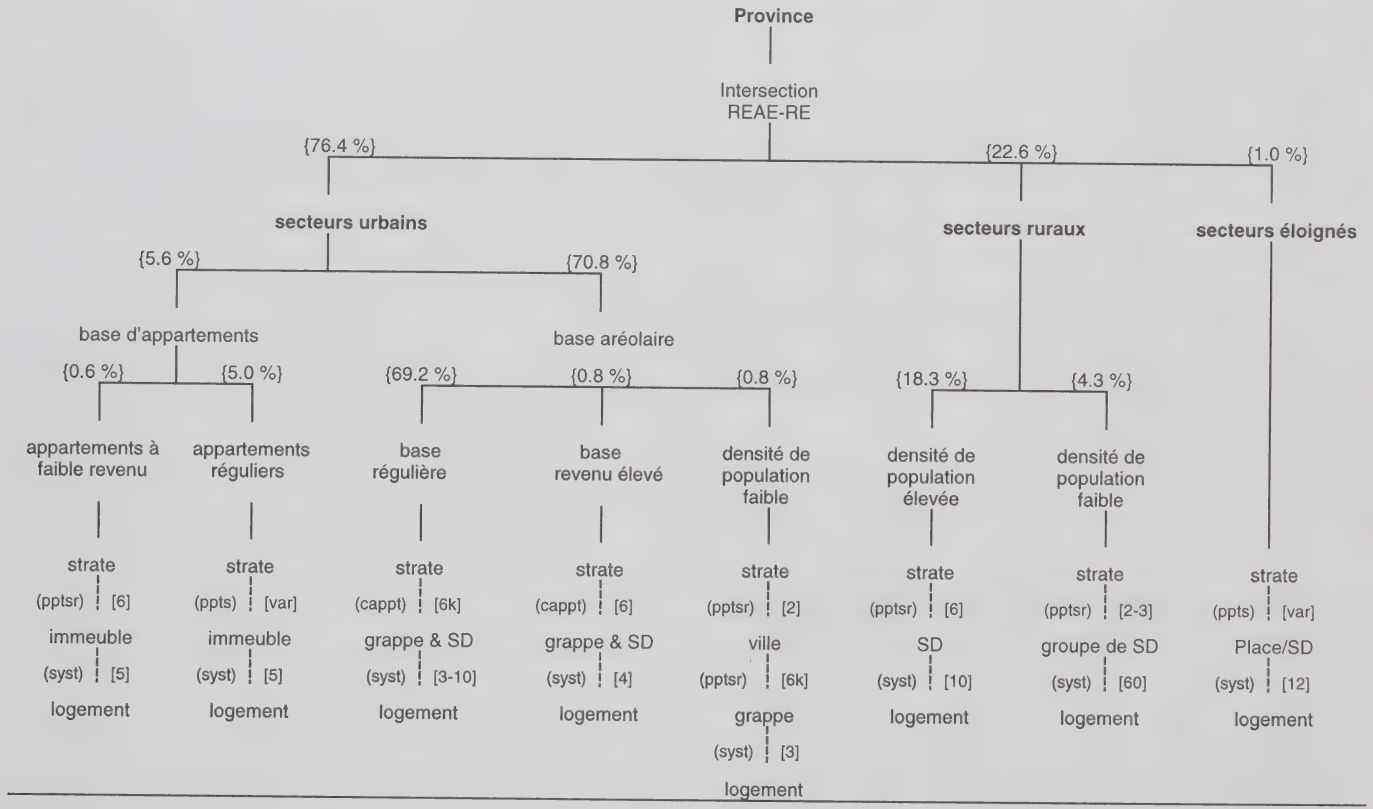
Manitoba -- plan FV: Dauphin, Flin Flon-the Pas, Swan River

Sasketchewan -- plan FV: Humboldt, Melfort, Melville, Nipawin

Colombie-Britannique -- plan SD: Hope, Parksville-Qualicum Beach, Trail-Rossland

Autres secteurs urbains: tous les autres secteurs urbains non mentionnés ci-dessus ne sont pas des strates. Ils sont soit des UPE, des portions d'une grande strate urbaine, ou ils sont incorporés dans une strate rurale.

Annexe C: Plan de sondage de l'Enquête sur la population active - 1995+



| = niveau de stratification

¦ = degré d'échantillonnage

REAE - Rég. économique d'assurance-emploi
RE - Région économique
{%} - pourcentage de l'échantillon total

SD - Secteur de dénombrement
grappe - ensemble de côtés d'îlots
( ) - méthode de sélection
[ ] - nombre d'unités sélectionnées
(6k = multiple de six,
var = nombre variable)

ppt - probabilité proportionnelle à la taille
ppts - ppt systématique
pptsr - ppts randomisé
cappt - classement aléatoire ppt
syst - systématique

Annexe B - **Abréviations**

AE	assurance-emploi
AR	agglomération de recensement
BR	bureau régional
CV	coefficient de variation
DR	division de recensement
DRHC	Développement des ressources humaines Canada
EPA	enquête sur la population active
ESCAPPT	(méthode d') échantillonnage systématique avec classement aléatoire et PPT
FRR	fichier de réseau routier
FSI	fraction de sondage inverse
FV	feuille de visites
PPCAO	programme de partage par circonscription assisté par ordinateur
PPT	probabilité proportionnelle à la taille
RAC	région d'assurance-chômage
RE	région économique
REAE	région économique d'assurance-emploi
RHC	Rao-Hartley-Cochran (méthode des groupes aléatoires)
RMR	région métropolitaine de recensement
SD	secteur de dénombrement
SDR	subdivision de recensement
SR	secteur de recensement
UPD	unité du premier degré
UPE	unité primaire d'échantillonnage

Tableau A3.2. Répartition de l'échantillon par région économique d'assurance-emploi, à compter de juillet 1995

Province	REAE	Taille.	Province	REAE	Taille	Province	REAE	Taille
Terre-Neuve	1	367	Québec (suite)	22	717	Ontario (suite)	42	760
	2	868		23	650		43	712
	3	649		24	571		44	786
Île-du-Prince-Édouard	4	1 421		25	598	Manitoba	45	1 920
Nouvelle-Écosse	5	439		26	538		46	1 056
	6	629	Ontario	27	847		47	837
	7	805		28	858	Saskatchewan	48	616
	8	702		29	685		49	623
	9	982		30	873		50	1 318
Nouveau-Brunswick	10	641		31	2 777		51	905
	11	592		32	738	Alberta	52	945
	12	1 222		33	752		53	1 057
	13	534		34	798		54	876
Québec	14	897		35	748		55	1 111
	15	778		36	539	Colombie-Britannique	56	748
	16	694		37	815		57	594
	17	564		38	815		58	1 757
	18	680		39	693		59	500
	19	723		40	849		60	610
	20	1 023		41	813		61	601
	21	2 134					Total = 52 350	

Tableau A3.1. Répartition de l'échantillon par région économique, à compter de juillet 1995

Province	RE	Taille	Province	RE	Taille	Province	RE	Taille
Terre-Neuve	10	680	Québec (suite)	450	553	Manitoba (suite)	660	273
	20	200		455	595		670	1 790
	30	473		460	685	Saskatchewan	710	957
	40	531		465	471		720	488
Île-du-Prince-Édouard	110	1 421		470	1 186		730	878
Nouvelle-Écosse	210	397		475	732		740	430
	220	630		480/490	404		750/760	709
	230	682	Ontario	510	1 455	Alberta	810	348
	240	1 002		515	926		820	300
	250	846		520	455		830	1 027
Nouveau-Brunswick	310	683		530	3 695		840	300
	320	576		540	2 193		850	300
	330	837		550	1 988		860	1 114
	340	519		560	1 102		870	300
	350	374		570	905		880	300
Québec	410	432		580	514	Colombie-Britannique	910	1 042
	415	432		590	1 740		920	2 423
	420	914		595	885		930	549
	425	462	Manitoba	610	320		940	200
	430	910		620	200		950	200
	435	1 387		630	374		970	200
	440	1 139		640	200		980	200
	445	265		650/680	657	Total = 52 350		

Tableau A1. Résultats de la stratification pour les strates à revenu élevé dans le cadre du remaniement de l'EPA

RMR	Logements	Strate	Grappes	Revenu médian ($)	Revenu moyen ($)
Montréal	15 237	3	83	121 881	132 818
Ottawa	6 558	2	39	111 729	116 973
Toronto	35 433	4	185	144 387	156 477
Hamilton	6 584	1	34	101 875	107 130
London	4 036	1	21	108 009	108 604
Winnipeg	7 543	2	42	96 763	100 264
Calgary	7 501	1	41	123 066	131 543
Edmonton	5 835	1	28	111 334	118 600
Vancouver	16 483	3	89	119 777	122 739
Total	105 210	18	562	122 765	132 217

Tableau A2. Taille de l'échantillon par province et CV dans le cadre de l'ancien plan, du remaniement et du plan actuel (après la réduction)

Province	Ancien plan				Remaniement				Plan actuel			
	Base		Total		Base		Total		Base		Total	
	Taille	CV	Taille	CV	Taille	CV	Taille	CV	Taille	CV	Taille	CV
Terre-Neuve	2 240	5.4	2 582	5.0	2 240	5.1	2 240	5.1	1 884	5.8	1 884	5.8
Île-du-Prince-Édouard	1 421	6.6	1 421	6.6	1 421	6.1	1 421	6.1	1 421	6.1	1 421	6.1
Nouvelle-Écosse	3 101	5.2	4 002	4.8	3 102	4.9	4 050	4.5	2 609	5.5	3 557	5.1
Nouveau-Brunswick	3 095	5.3	3 441	5.0	3 096	5.2	3 480	5.0	2 604	5.7	2 988	5.3
Québec	6 474	4.0	11 356	3.4	6 436	3.7	11 590	3.1	5 413	4.2	10 567	3.5
Ontario	8 517	3.8	17 388	3.2	8 473	3.4	17 206	2.8	7 125	3.7	15 858	2.9
Manitoba	3 276	6.3	3 897	6.1	3 869	5.0	4 428	4.8	3 254	5.6	3 813	5.3
Saskatchewan	4 527	5.0	4 563	5.0	4 053	5.0	4 107	5.0	3 409	5.6	3 463	5.6
Alberta	5 205	4.6	5 225	4.6	4 745	4.4	4 745	4.4	3 991	4.9	3 991	4.9
Colombie-Britannique	4 454	5.7	4 975	5.1	4 875	4.6	5 583	4.4	4 100	5.2	4 808	4.8
Canada	42 310	1.96	58 850	1.67	42 310	1.73	58 850	1.50	35 810	1.96	52 350	1.62

Tableau A0-2: Échantillon de l'EPA par REAE et autres subdivisions* -- 1998 (suite)

	Strate			Ménages				Source	
REAE	Urbain	Rurale	Appartement	Urbain	Rurale	Appartement	Total	Base	AE
546	18	0	5	684	0	191	875	571	304
547	16	9	0	553	527	0	1 080	387	693
548	29	14	0	901	561	0	1 462	657	805
549	18	0	2	780	0	94	874	132	742
550	66	0	16	1 971	0	827	2 798	2 798	0
551	14	0	4	604	0	133	737	448	289
552	17	0	1	760	0	56	816	258	558
553	15	0	2	641	0	157	798	282	516
554	7	4	0	290	185	0	475	136	339
555	11	0	2	558	0	81	638	167	471
556	17	0	2	720	0	95	815	238	577
557	5	2	0	353	227	0	580	262	318
558	20	8	0	843	529	0	1 372	439	933
559	17	0	0	760	0	0	760	73	687
560	15	0	0	712	0	0	712	52	660
561	13	3	0	587	129	0	716	279	437
664	45	0	6	1 622	0	273	1 895	1 895	0
665	10	11	0	522	742	0	1 264	1 139	125
666	4	3	0	337	410	0	747	221	526
767	18	0	0	767	0	0	767	562	205
768	16	0	1	709	0	63	773	623	150
769	14	17	0	581	1 023	0	1 604	1 604	0
770	8	4	0	414	514	0	928	619	309
871	25	0	3	861	0	85	946	946	0
872	29	0	3	970	0	89	1 059	1 059	0
873	21	13	0	1 072	914	0	1 986	1 986	0
975	12	4	0	530	218	0	748	748	0
976	60	0	6	1 573	0	196	1 769	1 769	0
977	13	0	1	612	0	38	649	351	298
978	21	4	0	628	356	0	984	627	357
979	8	4	0	539	280	0	819	618	201
Canada	954	210	68	36 454	13 627	2751	52 830	36 027	16 803

* Les subdivisions selon la base et AE sont approximatives

Annexe A

Tableau A0-1: Échantillon de l'EPA par province et autres subdivisions* -- 1998

	Strate			Ménages				Source	
Province	Urbaine	Rurale	Appartement	Urbain	Rurale	Appartement	Total	Base	AE
Terre-Neuve	35	13	0	1151	836	0	1 987	1 987	0
Île-du-Prince-Édouard	17	12	0	603	818	0	1 421	1 421	0
Nouvelle-Écosse	46	27	2	1753	1 714	90	3 557	2 611	946
Nouveau-Brunswick	45	20	0	1800	1 261	0	3 061	2 604	457
Québec	209	38	12	7693	2 383	283	10 358	5 458	4 900
Ontario	298	40	34	11 717	2 158	1 634	15 508	7 179	8 329
Manitoba	59	14	6	2 481	1 152	273	3 906	3 255	651
Saskatchewan	56	21	1	2 471	1 537	63	4 072	3 408	664
Alberta	75	13	6	2 903	914	174	3 991	3 991	0
Colombie-Britannique	114	12	7	3 882	854	234	4 969	4 113	856
Canada	954	210	68	36 454	13 627	2 751	52 830	36 027	16 803

Tableau A0-2: Échantillon de l'EPA par REAE et autres subdivisions* -- 1998

	Strate			Ménages				Source	
REAE	Urbain	Rurale	Appartement	Urbain	Rurale	Appartement	Total	Base	AE
020	15	0	0	465	0	0	465	465	0
021	20	13	0	686	836	0	1 522	1 522	0
123	17	12	0	603	818	0	1 421	1 421	0
224	8	2	0	294	145	0	439	397	42
225	7	6	0	287	342	0	629	555	74
226	22	0	2	716	0	90	806	806	0
227	4	7	0	235	466	0	701	490	211
228	5	12	0	221	761	0	982	363	619
329	32	4	0	1 270	225	0	1 495	1 090	405
330	13	16	0	530	1036	0	1 566	1 514	52
433	8	9	0	323	541	0	864	455	409
434	17	0	2	730	0	45	775	510	265
435	15	0	0	566	0	0	566	0	566
436	3	3	0	238	368	0	606	224	382
437	13	0	0	722	0	0	722	153	569
438	13	5	0	590	325	0	915	401	514
439	80	0	9	1 971	0	202	2 173	2 131	42
440	20	15	0	667	670	0	1 336	732	604
441	6	3	0	305	300	0	605	295	310
442	9	3	0	507	179	0	686	316	370
443	12	0	1	534	0	36	570	169	401
444	13	0	0	540	0	0	540	72	468

* Les subdivisions selon la base et AE sont approximatives

. . . suite . . .

Sunter, D., Kinack, M., Akyeampong, E. et Charette, D. (1995). Redesigning the Canadian Labour Force Survey Questionnaire: Development and Testing. Statistique Canada, document interne.

Tambay, J.-L. et Catlin, G. (1995). Plan d'échantillonnage de l'Enquête nationale sur la santé de la population. *Rapport sur la santé*, 7, 29-38.

Wolter K. M., (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Bibliographie

Brisebois, F. et Mantel, H. (1996). Month-in-sample effects for the Canadian Labour Force Survey. Congrès annuel de la SSC, juin 1996, *Recueil de la section des méthodes d'enquête.*

Brodeur, M., Montigny, G. et Bérard, H.(1995). Challenges in developing the National Longitudinal Survey of Children. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Cochran, W.G. (1977). *Sampling Techniques*, 3e Édition, John Wiley and Sons, New York.

Chen, E.J., Gambino, J., Laniel, N. et Lindeyer, J. (1994). Design and estimation issues for income in the redesign of the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Dufour, J., Simard, M., Allard, B. et Ray, G. (1996). Redesign of the Labour Force Survey Sample: impact on data quality. Statistique Canada, document interne.

Drew, J.D., Bélanger, Y. et Foy, P. (1985). La stratification dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 11, 109-124.

Friedman, H.P. et Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

Hartley, H.O. et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Kennedy B., Drew J. D., et Lorenz P. (1994). The Impact of Nonresponse Adjustment on Rotation Group Bias in the Canadian Labour Force Survey. Présenté au 5ième Atelier international sur la non-réponse dans les enquêtes auprès des ménages. Ottawa, Canada.

Lemaître, G.E. et Dufour J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête,* 13, 211-220.

Lorenz, P. (1995). Labour Force Survey -- Head Office Hot deck Imputation System Specifications -- Version 3. Statistique Canada, document interne.

Mantel, H., Laniel, N., Duval, M.-C. et Marion, J. (1994). Cost modelling of alternative sample designs for rural areas in the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Mian, I.U.H. et Laniel, J. (1994). Sample allocation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Rao, J.N.K., Hartley, H.O. et Cochran, W.G. (1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 24, 482-491.

Särndal, C.E., Swensson, B et Wretman J., (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Sheridan, M., Drew, D. et Allard, B. (1996). Le taux de réponse et l'Enquête sur la population active canadienne: Fruit de hasard ou bonne planification? Recueil de Statistique Canada Symposium 96 sur Erreurs non dues à l'échantillonage, 75-83.

Simard, M. et Dufour, J. (1995). Impact de l'implantation des interviews assistées par ordinateur comme nouvelle méthode de collecte à l'Enquête sur la population active. Statistique Canada, document interne.

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*. Statistique Canada. N° 71-526 au catalogue.

Singh, M.P., Gambino, J. et Mantel, H. (1994). Les petites régions: problèmes et solutions (avec discussion). *Techniques d'enquête,* 20, 3-23.

Singh, A.C., Kennedy, B., Wu, S. et Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Statistique Canada (1998). *Guide de l'Enquête sur la population active.* Disponible sur l'Internet à www.statcan.ca/francais/concepts/labour/index_f.htm

de détecter tout écart entre le nombre de logements recensés sur le terrain et le nombre de logements utilisés dans l'élaboration du plan de sondage. Le plan de sondage utilise un nombre qui est dérivé des dénombrements effectués à partir des données du recensement précédent. Par conséquent, toute divergence importante, c'est-à-dire tout écart de 50 % (positif ou négatif) entre le recensement et la dérivation, fait l'objet d'une enquête. Tout d'abord, toute grappe ayant un rendement inattendu est portée à l'attention de l'unité responsable du contrôle de l'échantillon à Ottawa, qui vérifie les frontières de la grappe et le nombre de logements attendu. Si l'écart ne peut être expliqué au bureau central, la grappe est acheminée au bureau régional concerné pour être analysée en détail. Toutes les causes expliquant les écarts sont répertoriées pour références futures.

Ce contrôle joue un rôle important puisque si la taille de l'échantillon nécessite des changements, il est essentiel de connaître quelles régions sont sous-échantillonnées ou suréchantillonnées. De plus, les écarts enregistrés peuvent révéler des problèmes pour l'enquête qui pourraient entacher la qualité des données de l'EPA.

Comités de l'EPA

L'EPA a besoin de plusieurs groupes de coordination pour veiller au bon déroulement de l'enquête. Certains comités jouent un rôle permanent à l'EPA, tandis que d'autres sont actifs uniquement pendant le remaniement. Durant le dernier remaniement, le Comité de direction du remaniement et le Comité des opérations du remaniement étaient les deux principaux comités. Le premier de ces comités, présidé par le directeur général de la direction des enquêtes des ménages et du travail, avait comme mandat de contrôler le remaniement dans son ensemble, c'est-à-dire de voir au déroulement en général des activités relatives à l'introduction de l'interview assistée par ordinateur, à la refonte de l'échantillon, au questionnaire remanié, aux nouveaux systèmes de traitement et aux nouveaux produits. Le Comité des opérations du remaniement, présidé par le directeur de la Division des enquêtes-ménages, avait comme mandat de gérer et de contrôler le travail sur ces projets en particulier.

Dans les lignes qui suivent, nous décrivons trois comités permanents. Leur mandat est de s'occuper de manière régulière des opérations permanentes et de l'évaluation de l'enquête.

Comité des opérations. Ce comité a pour mandat de passer en revue les activités survenues pendant les mois d'enquête et les circonstances entourant la tenue de l'enquête, de voir à ce que les opérations se déroulent bien, d'examiner les changements proposés et d'en recommander l'adoption ; le tout afin que l'enquête continue d'atteindre ses objectifs. Ce comité, qui se réunit toutes les deux semaines, est présidé par un membre de la Division des enquêtes-ménages.

Comité de direction des estimations de population. Ce comité a pour mandat de contrôler les estimations de population postcensitaires requises par l'EPA. Ce comité évalue également les sources de données utilisées et les méthodes appliquées pour obtenir les estimations à différents niveaux géographiques, et met en oeuvre plusieurs projets de recherche sur le sujet. Ce comité est présidé par un membre de la Division de l'analyse des enquêtes sur le travail et les ménages.

Comité sur la qualité des données. Le comité actuel, qui a officiellement vu le jour au printemps de 1972, avait à l'époque comme fonction de diffuser la qualité des données de l'EPA et de ses suppléments. Depuis ce temps, le mandat de ce comité s'est quelque peu élargi. Son mandat consiste maintenant à examiner et à évaluer la qualité des données de l'EPA sur une base mensuelle, à proposer et réviser des projets de recherche et de développement visant la mise au point de méthodes pouvant influer sur la qualité des données, et enfin à surveiller la recherche et le développement dans ce domaine. Ce comité est présidé par un membre de la Division des méthodes d'enquêtes auprès des ménages.

Pour assurer la meilleure qualité de données possible, le Comité sur la qualité des données examine périodiquement les différents indicateurs de qualité décrits plus haut. Le comité se réunit chaque mois pour étudier et évaluer la qualité des données mensuelles et pour faire des suggestions et des recommandations sur tout aspect susceptible d'amélioration. Les membres du comité disposent d'un certain nombre de documents pour mener à bien leur tâche. En suivant étroitement l'évolution des indicateurs, le comité peut intervenir immédiatement auprès des responsables des activités de l'EPA concernées afin de contrôler la qualité des données mensuelles. Le comité discute également de nouveaux faits susceptibles d'influencer la qualité des données venant d'être recueillies ou devant être recueillies dans l'avenir, en particulier les changements relatifs aux méthodes de collecte ou au questionnaire, des problèmes inhabituels sur le terrain, de la mise à l'essai continue des procédés et des méthodes, etc.

taux de rejet à la vérification et aux taux de réponse. Les intervieweurs principaux sont régulièrement en contact avec leurs intervieweurs et ils portent à leur attention les résultats des divers indicateurs de rendement.

Vérification des logements codés vacants. Le programme de vérification des logements codés vacants a pour but de contrôler le travail effectué par les intervieweurs sur le terrain. Pour chacun des intervieweurs, au moins une fois tous les vingt-quatre mois, un échantillon de logements codés vacants est sélectionné. L'intervieweur principal retourne à ces logements pour vérifier si les logements étaient effectivement vacants et les logements sont recodés (vacants ou autre). Suite aux résultats de ce programme, l'intervieweur recevra de la formation additionnelle si on le juge nécessaire. Cette information sert également à mesurer combien de ménages codés hors du champ de l'enquête ont contribué positivement au glissement, bien qu'il soit très difficile d'extrapoler pour l'échantillon complet puisque le choix des logements vérifiés est laissé à la discrétion des BR.

Programme de validation. Le programme de validation a été conçu afin de contrôler la performance des intervieweurs et de fournir aux intervieweurs une rétroaction rectificative sous forme de formation additionnelle selon les faiblesses identifiées. Les intervieweurs sont soumis à la validation de façon aléatoire de sorte que, durant une année, chacun d'entre eux est choisi deux fois. Environ 2 % des ménages font partie de ce programme chaque mois (à l'exception des mois d'avril et de décembre où ce programme n'a pas lieu). La semaine suivant la semaine d'enquête, les intervieweurs principaux recontactent les personnes qui ont fourni l'information durant la semaine d'enquête pour l'échantillon soumis à ce programme. Ils leurs posent des questions de politesse telles que : confirmation de l'adresse, souvenir d'avoir participé à l'enquête, moment de l'interview, attitude de l'intervieweur, etc. L'intervieweur principal profite également de l'occasion pour remercier le répondant pour sa participation et son temps. Ce programme a été suspendu depuis l'introduction du mode de collecte d'interview assistée par ordinateur, mais sera rétabli bientôt sous forme informatisée.

Réinterview. Le programme de réinterview de l'EPA a pour but de mesurer la variabilité de réponse associée aux données recueillies par l'enquête et de déterminer les causes des erreurs de réponse ainsi que les conditions dans lesquelles ces erreurs sont susceptibles de se produire. En plus des erreurs attribuables au répondant et à l'intervieweur, le terme « erreur de réponse » comprend les erreurs dues aux imperfections dans le questionnaire, par exemple aux maladresses de formulation des questions ou à l'imprécision du codage. Ce programme a été suspendu au début des années 1990 et, pour l'instant, son avenir est incertain.

L'analyse périodique des données obtenues de ce programme permettait de découvrir les causes d'erreur et d'améliorer continuellement l'ensemble des méthodes utilisées, puisqu'il y avait des incidences sur divers éléments de la conception d'une enquête tels que la composition du questionnaire, les techniques d'interview, la formation des intervieweurs et d'autres encore. Dans ce qui suit, on explique en quoi consistait ce programme.

Tous les mois, sauf en décembre, on choisissait un certain nombre de logements faisant partie de l'échantillon de l'EPA pour le programme de réinterview. Ce sous-échantillon était prélevé de manière à ce qu'une partie de la tâche de chaque intervieweur soit réinterviewée au moins deux fois par an. Les surveillants au BR, ou les intervieweurs principaux, procédaient à la réinterview par téléphone dans la semaine suivant la semaine d'enquête, en tenant compte du fait que la semaine de référence avait lieu deux semaines auparavant.

L'échantillon constitué pour la réinterview était divisé en deux parties. Pour l'une d'entre elles, l'intervieweur principal comparait les réponses obtenues à la seconde interview avec celles fournies la première fois à l'intervieweur et s'il y avait des différences, déterminait avec l'aide du répondant quelle était la bonne réponse. Pour l'autre partie de l'échantillon, l'intervieweur principal se contentait de mener une interview indépendante. On se servait des données recueillies auprès de la partie de l'échantillon qui avait fait l'objet de la comparaison pour obtenir une mesure du biais de réponse ; l'hypothèse étant que l'intervieweur et l'intervieweur principal n'auraient pas fait les mêmes erreurs et que la comparaison permettrait de découvrir les bonnes réponses. On se servait des données recueillies auprès de la partie de l'échantillon qui n'avait pas fait l'objet de la comparaison pour estimer la variance de la réponse. La comparaison permettait aussi à l'intervieweur principal de vérifier le travail des intervieweurs. Après la réinterview, l'intervieweur principal discutait avec les intervieweurs des divergences obtenues dans les résultats, il leur donnait des conseils et leur signalait les questions suscitant le plus grand nombre d'erreurs.

Contrôle du rendement des grappes. Le rendement des grappes de l'EPA est contrôlé sur une base mensuelle afin

opérations, mais cette fois on s'intéresse à l'échelle provinciale plutôt que régionale. Contrairement au Rapport sur les opérations, le Rapport de qualité renferme des textes qui commentent les différents tableaux et graphiques. À ces indicateurs réguliers, s'ajoute à l'occasion un chapitre spécial sur un sujet d'intérêt particulier.

Mise à jour des coefficients de variation. Conformément à la politique sur les critères de qualité pour la diffusion à Statistique Canada, l'EPA produit chaque mois les variances correspondant aux estimations produites. Les CV moyens sur six mois sont mis à jour deux fois par année.

Rapports spéciaux. En plus des rapports produits régulièrement pour assurer et contrôler la qualité des données de l'EPA, certains rapports spéciaux sont rédigés à l'occasion. Un exemple de ce rapport est le rapport sur l'impact de l'introduction de l'interview assistée par ordinateur sur la qualité des données de l'EPA (Simard et Dufour, 1995). Puisque cette introduction s'est faite d'une façon graduelle, il a été ainsi possible de comparer la portion de l'échantillon interviewée à l'aide de l'interview assistée par ordinateur avec celle sondée au moyen de la méthode papier-crayon, et ce, selon plusieurs indicateurs de qualité. Le nombre de données intéressantes sur le sujet a fait naître un rapport spécial substantiel. Ce fut également le cas pour le rapport sur l'impact de la refonte du plan d'échantillonnage de l'EPA sur la qualité des données (voir Dufour, Simard, Allard et Ray; 1996).

Programmes d'assurance de la qualité

Au fil des ans, l'EPA s'est munie de plusieurs programmes pour assurer la qualité des données qu'elle diffuse. Certains d'entre eux ont cependant dû être délaissés en raison des compressions budgétaires. Huit de ces programmes sont présentés dans les lignes qui suivent.

Recrutement. Avant d'embaucher des candidats à des postes d'intervieweurs, on évalue leurs aptitudes et leur capacité à bien remplir les documents d'enquête. Avant même que la formation commence, on leur envoie un exemplaire de plusieurs documents. Ces documents décrivent le travail des intervieweurs à Statistique Canada (en mentionnant les responsabilités, les techniques et les compétences requises), l'organisation de Statistique Canada et l'ordinateur portatif comme instrument de collecte.

Formation. La période initiale de formation des intervieweurs dure deux mois. Elle commence par un cours de trois jours en classe, pendant lequel on montre aux intervieweurs qui viennent d'être embauchés comment utiliser leur ordinateur portatif et l'équipement informatique, comment remplir les formulaires d'enquête et les documents administratifs. En outre, ces derniers y font des exercices pratiques, simulent des interviews et apprennent les techniques d'interview.

Ensuite les intervieweurs reçoivent deux jours de formation en milieu de travail pendant leur première semaine d'enquête et un jour ou deux pendant la deuxième semaine, au besoin. Pendant ce temps, un intervieweur principal les accompagne, les observe, leur indique comment mener les interviews et leur donne l'exemple en effectuant lui-même des interviews. Tous les intervieweurs participent également à des séances spéciales de formation et de recyclage en groupe, au moins une fois par an.

Le travail des intervieweurs est évalué dans le cadre d'autres programmes, qui seront décrits plus loin. Selon le rendement de chacun, on détermine s'il y a lieu d'ajouter des cours d'autoformation ou des exercices de révision pour éclaircir certains points ou remédier à des faiblesses.

Observation. Le programme d'observation vise à réduire au maximum les erreurs que les intervieweurs pourraient commettre en donnant à l'intervieweur principal l'occasion d'observer ceux qui relèvent de lui, d'évaluer leur rendement et de déceler les problèmes. Chaque intervieweur est observé au moins une fois tous les vingt-quatre mois. C'est au bureau régional qu'on décide qui sera observé et quand, de manière à ce qu'on ne puisse pas deviner l'ordre dans lequel le programme d'observation se déroule. En dehors de ce programme, il est possible pour l'intervieweur principal d'observer un de ses intervieweurs s'il soupçonne un problème en particulier. L'intervieweur principal accompagne l'intervieweur pendant toute une journée et voit comment se passent les interviews sur place et les interviews téléphoniques. La deuxième journée, l'intervieweur principal vérifie les listes de grappe. Il envoie ensuite les résultats de l'observation au BR et rédige des rapports périodiques à l'intention du bureau central. L'intervieweur principal transmet à l'intervieweur le résultat de sa performance aussitôt que possible après l'observation.

Rétroaction sur le rendement. À tous les mois, le rendement des intervieweurs fait l'objet de rapports mensuels aux BR. Ces rapports ont trait aux coûts, aux

rejoindre un répondant, les meilleurs moments (heure et jour) pour réaliser une interview, etc. On peut également vérifier si les procédures de collecte sont suivies à la lettre par les intervieweurs. Par exemple, des codes de non-réponse signifiant « personne à la maison » ne doivent être utilisés qu'à la fin de la semaine d'enquête seulement. Ce nouveau système permet donc d'en connaître davantage sur le travail exécuté sur le terrain, de réagir en présence de cas douteux (comme par exemple des interviews effectuées en moins d'une minute) et, par conséquent, de réduire certaines erreurs qui peuvent se glisser. Cette nouvelle information peut également être utilisée pour améliorer le programme de formation donné aux intervieweurs et renforcer certains comportements tels que la planification des tâches ou l'horaire de travail. Cette base de données permet également de connaître le nombre de cas qui sont convertis de refus à répondant, ce qui n'était pas possible auparavant. En fait, le système de gestion de cas permet de connaître toutes les actions qui ont été entreprises sur un dossier. Un groupe de travail est actuellement à l'étape de développer un rapport opérationnel dont le but sera d'informer les bureaux régionaux sur toute situation jugée inhabituelle enregistrée durant le mois d'enquête (ex. : interviews effectuées avant la semaine d'enquête, période trop courte entre deux interviews sur place, appels nocturnes, etc.). Pour être efficace, ce rapport doit être produit dans les plus brefs délais suivant la semaine d'enquête pour pallier l'effet mémoire lorsque l'on cherche à en connaître la cause en contactant les intervieweurs.

Les nouvelles données du système de gestion de cas permettent également d'en savoir plus sur les règles de vérification effectuées au moment de l'entrevue. Avec le nouveau système, il est par exemple possible de savoir combien de règles de vérification ont été outrepassées après confirmation avec le répondant, combien de fois une règle de vérification a été appliquée et le nombre de fois qu'une observation a échoué aux règles.

Avec l'introduction du nouveau questionnaire et du nouveau système informatique, il est maintenant possible de produire un taux d'imputation global, par questionnaire et par question ; ce qui était impossible auparavant.

Toutes ces nouvelles mesures permettront de mieux comprendre ce qui se passe sur le terrain, et aideront lors de l'interprétation des résultats.

Rapports sur la qualité des données de l'EPA

Un certain nombre de rapports sont mis à la disposition des utilisateurs des données de l'EPA au bureau central à Ottawa ou dans les régions. De plus, ces documents sont consultés de façon régulière par les membres du Comité sur la qualité des données de l'EPA pour s'assurer de la qualité de l'enquête. En général, ils contiennent toute une gamme d'indicateurs de qualité à des niveaux géographiques différents et pour des périodes de temps plus ou moins longues.

Rapport mensuel sur les opérations de l'enquête de l'EPA. Tous les mois, l'unité de la qualité des données de l'EPA produit un rapport sur la qualité des données de l'enquête du mois courant. Ce rapport est par ailleurs discuté la veille de la diffusion des données de l'enquête. Son but principal est de permettre de contrôler la qualité des opérations sur le terrain, c'est pourquoi la plupart des indicateurs de qualité y sont présentés par bureau régional. Certaines séries sont également présentées pour une période couvrant 26 mois pour mieux saisir les changements saisonniers et les changements mensuels en comparaison avec une année précédente. Le rapport contient les indicateurs de qualité suivants : taux de non-réponse (par BR, composante, nombre de mois dans l'enquête, type de secteur), taux de vacance (par BR, type de secteur), taux de glissement (par province et groupe d'âge-sexe), taille de l'échantillon, nombre de problèmes techniques et nombre de dossiers temporaires.

Tableaux de variance. Les tableaux de variance, produits chaque mois, contiennent les estimations, les CV, les variances et les effets du plan des principales caractéristiques de l'EPA à différents niveaux géographiques. À ces indicateurs, se joignent la taille moyenne des ménages et les taux de glissement à des niveaux plus détaillés que ceux présentés dans le Rapport sur les opérations de l'EPA. À l'heure actuelle, on songe à adopter un format exclusivement électronique, plus moderne.

Rapport de qualité. Le Rapport de qualité de l'EPA est produit deux fois par année et couvre les périodes de janvier à juin et de juillet à décembre. Le but de ce rapport est de présenter un examen approfondi des mesures de la qualité associées à l'EPA pour la période de six mois à l'étude. De plus, on analyse les mesures de la qualité sur une période de trente mois dans le but de déceler des tendances ou les effets de certains changements apportés aux activités ou au plan de sondage. Dans ce document, on reprend plusieurs indicateurs de la qualité présentés dans le Rapport sur les

groupe de renouvellement est rajusté selon la population totale de la province. Enfin, le rajustement pour la non-réponse est maintenant effectué séparément par groupe de renouvellement, à la différence de la période étudiée par Brisebois et Mantel. Il en résulte que l'effet des écarts entre groupes de renouvellement selon ces poids finaux est probablement encore plus faible.

Erreurs de traitement

Les erreurs de traitement peuvent se produire à diverses étapes de l'enquête, soit à la saisie, à la vérification, au codage, à la pondération ou à la totalisation des données.

L'utilisation d'un mode de collecte informatisé permet d'éviter des erreurs d'aiguillage dans le questionnaire, puisque c'est maintenant l'application qui détermine la prochaine question à poser compte tenu des réponses préalablement entrées. De même, certaines règles de vérification sont incorporées au système de collecte, ce qui permet de détecter et de corriger certaines divergences au moment de l'interview. Toutefois, toutes les règles de vérification n'ont pu être incorporées à l'instrument de collecte informatisé, puisqu'il fallait faire un compromis entre la durée de l'interview, la vitesse de l'ordinateur et l'efficacité de celui-ci. Le bureau central complète donc cette étape en appliquant un ensemble de règles de vérification en lot.

Le module de contrôle sur le terrain fournit deux types d'indicateurs de qualité : le taux d'échec et le taux de divergence au contrôle des formulaires. Le taux d'échec correspond au pourcentage de questionnaires qui comportaient au moins une divergence. Une divergence est définie comme toute inscription effacée, modifiée ou ajoutée à une zone en blanc après soumission à certains contrôles pour en vérifier la validité. Le taux de divergence au contrôle, quant à lui, représente le pourcentage de divergences sur un questionnaire par rapport au nombre total d'entrées sur le questionnaire. On relève des taux d'échec respectifs d'environ de 1 et 5 % pour les caractéristiques démographiques et d'autres aspects. Les taux de divergence au contrôle correspondants s'élèvent respectivement à 0,1 et 1 %.

Les codages automatique et manuel des variables profession et branche d'activité sont effectués au bureau central. Au premier mois d'interview ou en présence de tout changement pour ces deux variables, l'intervieweur recueille de l'information décrivant avec précision le genre d'entreprise, d'industrie ou de service où la personne travaille, et de l'information indiquant clairement et précisément le genre de travail ou la nature des fonctions. Le premier type d'information servira à déterminer la branche d'activité, tandis que le second permettra d'identifier la profession. Une des premières étapes du traitement au bureau central consiste à coder automatiquement l'information descriptive recueillie pour les variables profession et branche d'activité selon la classification standard pour ces variables, soit la CTI et CTP. Les enregistrements ne pouvant être codés par le système automatique sont codés manuellement par une équipe de codeurs de l'EPA. Environ 14 000 enregistrements sont codés à la main chaque mois. Afin de contrôler la qualité du codage effectué manuellement, un plan statistique de contrôle de la qualité est appliqué mensuellement. Les trois mesures utilisées pour connaître l'efficacité de ce processus de contrôle sont le taux de vérification, qui est le pourcentage d'enregistrements vérifiés sur l'ensemble des enregistrements soumis au contrôle, la qualité moyenne à l'entrée (QME), qui est une estimation du pourcentage d'enregistrements contenant des erreurs avant d'appliquer le contrôle de la qualité, et la qualité moyenne à la sortie (QMS), qui est une estimation du pourcentage d'enregistrements contenant encore des erreurs de codage après le contrôle de la qualité. En 1995, la valeur moyenne de ces trois mesures était de 19,6 % pour le taux de vérification, 7,9 % pour la QME et de 4,8 % pour la QMS.

Pour éviter les erreurs susceptibles de se produire à l'étape de l'estimation et de la totalisation, on procède à un examen détaillé du résultat de ces activités, à une analyse de différents diagnostics produits automatiquement par le système et à une comparaison avec d'autres sources de données.

Nouvelles mesures d'erreur

L'adoption du mode de collecte informatisé à l'automne 1993 a généré toute une gamme de données qui étaient difficilement accessibles auparavant. En fait, le mode d'interview assistée par ordinateur est doté d'un système de gestion de cas dont les fonctions principales consistent à acheminer les cas, à reporter et à assister les intervieweurs durant le déroulement de l'enquête. Toutes les actions faites sur un cas sont dorénavant enregistrées. Conséquemment, au cours de chaque mois de l'enquête, des fichiers sont produits directement par le système de gestion des cas et contiennent une foule de renseignements sur ce qui se passe sur le terrain. Par exemple, on peut obtenir le temps moyen par interview personnelle ou téléphonique, le nombre de tentatives pour

simplement retirés de la base de sondage. Une attention particulière doit être apportée à la détermination des logements qui sont vacants puisque ceux-ci influencent directement deux autres indicateurs. En effet, si un logement est codé vacant plutôt que d'être identifié comme étant du type temporairement absent par exemple, le taux de non-réponse produit pour l'EPA est quelque peu sous-estimé. Par ailleur, le taux de glissement s'en trouve surestimé puisque ce logement mal codé aurait dû être considéré lors de la détermination de ce taux. Les intervieweurs se doivent donc de faire un travail très minutieux pour déterminer si un logement est vacant, et par conséquent, hors du champ de l'enquête, ou tout simplement occupé par un ménage temporairement absent et donc dans le champ de l'enquête.

Le tableau suivant présente les taux de vacance moyens, les valeurs minimales et maximales pour l'année 1997 à l'échelle provinciale et nationale. Encore une fois, pour répondre aux normes et lignes directrices de Statistique Canada pour la déclaration de la non-réponse, des taux pondérés et non pondérés sont produits mensuellement et acheminés à la banque de données centrale sur la non-réponse de Statistique Canada. En général, le taux de vacance est relativement stable, affichant une tendance à la hausse plus on s'éloigne du dernier remaniement puisque la base de sondage est moins à jour. Après chaque remaniement, le taux de vacance affiche une tendance à la baisse. Cette baisse a été d'autant plus marquée après le dernier remaniement étant donné le caractère plus urbanisé du nouveau plan de sondage. Pour cet indicateur de qualité, certaines provinces se détachent des autres en affichant des taux beaucoup plus bas ou plus élevés.

Taux de vacance (non pondéré), Canada et provinces - 1997

Province	Moyenne (%)	Min (%)	Max (%)
Terre-Neuve	15,4	14,9	16,4
Île-du-Prince-Édouard	20,5	18,6	23,0
Nouvelle-Écosse	16,8	15,2	18,7
Nouveau-Brunswick	14,1	13,5	15,2
Québec	14,0	11,9	15,8
Ontario	10,8	10,0	11,3
Manitoba	17,1	16,4	17,7
Saskatchewan	14,7	12,5	15,5
Alberta	8,7	8,1	9,8
Colombie-Britannique	9,5	8,7	9,8
Canada	13,0	12,2	13,5

Erreur de réponse

Cette erreur peut être attribuable à la conception du questionnaire, à la formulation des questions, à la compréhension du répondant, à la façon dont l'interview est menée ainsi qu'aux conditions générales dans lesquelles l'enquête est réalisée. Des erreurs de réponse peuvent se produire au moment où les renseignements sont fournis, reçus ou entrés sur l'ordinateur portatif. Toutefois, le mode de collecte informatisé permet de réduire certaines de ces erreurs, puisque maintenant certaines règles de vérification sont incorporées à l'instrument de collecte et les conflits doivent être résolus au moment même de l'interview. Il se peut toutefois que le répondant interprète mal la question, qu'il ne sache pas la réponse, qu'il ait oublié ou qu'il préfère déformer les faits pour des raisons qui lui sont personnelles. De plus, il arrive que les intervieweurs aient tendance à expliquer les réponses ou à les interpréter de manière différente. Les erreurs de réponse, comme les autres catégories d'erreurs, peuvent avoir une variance et un biais.

Les réponses par procuration (ou par personne interposée), que l'on obtient lorsqu'on recueille auprès d'un membre du ménage des renseignements se rapportant à un autre membre, peuvent aussi entraîner des erreurs de réponse.

Dans les enquêtes répétées, où l'échantillon est constitué d'un certain nombre de panels ou de groupes de renouvellement, l'espérance mathématique des estimations varie légèrement d'un groupe de renouvellement à un autre. Il se produit alors ce qu'on appelle un biais de renouvellement. En ce qui concerne l'EPA, ce biais atteint son plus haut niveau pour le sixième de l'échantillon qui en est à sa première interview. On peut obtenir l'indice de renouvellement en faisant le rapport entre une estimation calculée pour la partie de l'échantillon participant à l'enquête pour un certain nombre de fois (premier mois, deuxième, etc.) et l'estimation calculée pour l'échantillon entier.

Brisebois et Mantel (1996) ont calculé un indice de renouvellement modifié qui tient compte des différences des effets des erreurs dues à l'échantillonnage pour les six groupes de renouvellement. Leur étude, qui se base sur un échantillon pondéré avant rajustement pour contrôles démographiques, a permis de relever plusieurs différences statistiquement significatives parmi les groupes de renouvellement mais dont l'effet concret est mineur. En pratique, les estimations publiées sont basées sur des pondérations ajustées selon l'âge-sexe et les contrôles de population géographique. En outre, le poids de chaque

classifiés vacants par erreur. À l'EPA, un programme intitulé Programme de vérification des logements vacants a été mis sur pied pour obtenir de l'information sur cette erreur.

Depuis 1993, l'EPA se soumet aux normes et lignes directrices de Statistique Canada pour la déclaration des taux de non-réponse. Tous les mois, les taux de non-réponse pondérés et non pondérés sont acheminés à la banque de données centrale sur la non-réponse de Statistique Canada, dont le mandat est de compiler les données longitudinales pour plusieurs enquêtes régulières. Cette base de données exige des taux de non-réponse à l'étape de la collecte et à l'étape de l'estimation. Avant le remaniement, l'EPA ne fournissait des taux que pour l'étape de la collecte car ils étaient les mêmes qu'à l'étape de l'estimation. Avec la mise en application du nouveau questionnaire et des nouveaux systèmes de production en 1997, il est maintenant possible de produire des taux différents pour les étapes de la collecte et de l'estimation.

Le tableau suivant présente les taux de non-réponse moyens pour 1'année 1997 ainsi que le minimum et le maximum atteints durant cette année. À l'EPA, le maximum pour la non-réponse est normalement atteint au mois de juillet, étant donné le haut pourcentage de personnes qui ne sont pas à la maison, et le minimum au mois d'octobre. Depuis la fin de l'année 1993, plusieurs facteurs ont perturbé la série du taux de non-réponse à 1'EPA. Tout d'abord, l' introduction de l'interview assistée par ordinateur a généré un nouveau type de non-réponse qui était pratiquement inexistant auparavant. L'urbanisation du plan de sondage (introduit progressivement du mois d'octobre 1994 au mois de février 1995), c'est-à-dire une plus grande proportion de l'échantillon provenant des secteurs urbains, a également eu un effet, quoique négligeable, sur cette série puisqu'on obtient en général des taux de non-réponse plus élevés en région urbaine qu'en région rurale. Finalement, le nouveau plan d'échantillonnage a nécessité l'embauche de nouveaux intervieweurs qui ont tendance a obtenir des taux de non-réponse légèrement plus élevés durant leurs six premiers mois à l'EPA. Pour une revue historique des enjeux de la non-réponse à l'EPA, se référer à l'article de Sheridan et coll. (1996).

Taux de non-réponse (non pondéré), Canada et provinces - 1997

Province	Moyenne (%)	Max (%)	Min (%)
Terre-Neuve	4,2	5,4	3,0
Île-du-Prince-Édouard	3,5	4,8	2,4
Nouvelle-Écosse	6,3	7,3	4,6
Nouveau-Brunswick	4,6	5,4	3,1
Québec	5,4	6,6	3,7
Ontario	4,8	5,7	3,7
Manitoba	3,6	5,4	2,1
Saskatchewan	3,6	4,6	2,4
Alberta	4,9	6,3	3,1
Colombie-Britannique	5,7	6,7	4,5
Canada	4,9	5,5	3,8

Durant les essais qui ont précédé la mise en place de la nouvelle méthode de collecte assistée par ordinateur, les gestionnaires de l'EPA étaient préoccupés par les effets possibles de ces changements sur les refus de participer à l'enquête, étant donné la présence d'un ordinateur portatif au domicile du répondant lors de la première entrevue susceptible de susciter plus de réticence de la part du répondant. Conséquemment, une attention particulière a été portée à cette composante de la non-réponse durant les essais, et aucune augmentation importante, qui puisse être directement liée à l'interview assistée par ordinateur, n'a été observée.

Les taux de refus pour l'EPA sont habituellement très bas. Les taux mensuels canadiens varient entre 1 et 2 %. Les taux de refus à l'échelle provinciale sont ordinairement du même ordre de grandeur, mais ils peuvent descendre aussi bas que 0,5 % ou monter aussi haut que 3 %. Par ailleurs, l'avènement du mode de collecte informatisé a permis, pour la première fois de l'histoire de l'EPA, d'appliquer des grilles automatisées mentionnant les motifs de refus de participation à cette enquête comme par exemple : le manque de temps, ne croit pas aux statistiques, trop personnel, contre le gouvernement. Des études peuvent donc maintenant être entreprises plus facilement pour déceler les causes de refus et tenter de les réduire au maximum. On peut également obtenir du système de gestion de cas, propre au mode de collecte informatisé, le nombre de refus qui sont convertis en répondants par les intervieweurs principaux à tous les mois. Cette donnée était autrefois inexistante.

Vacance

Les logements identifiés correctement comme étant vacants ou inexistants n'introduisent aucun biais dans les estimations de 1 'EPA. Par contre, la variance de l'estimation s'en trouve plus élevée puisque l'échantillon compte un nombre moins élevé de ménages. Les intervieweurs de l'EPA retournent visiter les logements vacants à tous les mois afin d'interviewer les personnes ciblées par l'enquête qui peuvent s'y être installées depuis l'enquête précédente. Les logements inexistants sont tout

admissibles à l'enquête. Les logements identifiés non admissibles pour le mois d'enquête le sont pour les raisons suivantes :

- logement hors du champ d'enquête, c'est-à-dire un logement occupé par des personnes ne faisant pas partie de la population cible, p. ex., des membres des Forces armées canadiennes ;

- logement vacant : logement non occupé, logement saisonnier ou logement en construction ;

- logement non existant : logement démoli, logement transformé en local d'affaires, maison mobile déménagée ou encore logement abandonné ou inscrit par erreur.

Lorsqu'un logement a été identifié comme étant admissible à l'enquête, il n'est pas toujours possible de réaliser l'interview, et ce pour les raisons suivantes :

– non-réponse du ménage : personne à la maison, absence temporaire, interview impossible (mauvais temps, circonstances inhabituelles dans le ménage, etc.) ou refus.

De plus, depuis l'introduction de l'interview assistée par ordinateur à l'automne 1993, un nouveau code de non-réponse a fait son apparition. Ce code, qui auparavant représentait les questionnaires non reçus à temps pour le traitement à cause de problèmes postaux, garde aujourd'hui la même connotation mais est attribuable à des problèmes d'ordre technique. Parmi ces problèmes figurent des situations telles : disque dur en panne, défectuosité du système d'entraînement des bandes magnétiques, allocation de mémoire insuffisante, surcharge de chaleur, pannes de courant, ennuis téléphoniques, etc. La majorité de ces cas peuvent être résolus pour le mois suivant, mais étant donné les délais de publication très courts, il est souvent impossible de les résoudre durant le mois courant. En conséquence, ces cas sont considérés comme non-répondants. Cette composante de la non-réponse est devenue presque négligeable à mesure que s'est accrue notre expérience de l'interview assistée par ordinateur.

On compense pour les unités non répondantes en faisant appel à l'une des trois approches suivantes. Les ménages non répondants pour le mois courant se verront attribuer l'information fournie le mois précédent, si elle existe. Cette procédure ne peut toutefois pas être appliquée deux mois consécutifs, et permet de traiter environ 30 % des non-répondants. Pour les ménages non répondants, dont on ne connaît aucune information provenant du mois précédent, on compense en gonflant le poids des ménages répondants qui appartiennent au même groupe de renouvellement, à la même région économique d'assurance-emploi et au même type de secteur (voir Chapitre 5), par un facteur équivalant à l'inverse du taux de réponse. L'importance du biais attribuable à la non-réponse est habituellement inconnue, mais on sait qu'elle est directement liée aux différences de caractéristiques entre les groupes d'unités répondantes et les groupes d'unités non répondantes. C'est d'ailleurs une des raisons pour laquelle on a ajouté la variable groupe de renouvellement comme variable de regroupement au moment d'effectuer une compensation pour la non-réponse. Plusieurs études ont démontré que la non-réponse affiche un comportement différent selon la durée de participation à l'enquête. Comme l'effet de ce biais s'accroît avec l'augmentation du taux de non-réponse, on essaie de maintenir le taux de réponse à un niveau aussi élevé que possible durant les activités de collecte.

Pour la non-réponse partielle, on a recours à une méthode d'imputation. Dans un premier temps, on applique si possible l'imputation déterministe, c'est-à-dire qu'on procède à l'examen des réponses recueillies aux autres questions qui s'y rattachent et une seule valeur est jugée possible. En cas d'échec, on fait appel à une méthode d'imputation dite « hot deck ». Le nouveau système d'imputation choisit un donneur au hasard parmi les enregistrements du mois courant ayant passé les règles de vérification. Si le processus itératif de recherche d'un donneur échoue ou si l'enregistrement imputé ne satisfait pas aux règles de vérification étant donné la non-concordance entre les données recueillies et celles imputées après un certain nombre de tentatives, l'enregistrement en défaut est totalement substitué. Pour la variable des mesures de gains, qui a été introduite en 1997 avec le nouveau questionnaire, on utilise plutôt une méthode d'imputation dite « warm deck ». En effet, pour cette variable, le choix des donneurs n'est pas restreint uniquement aux données du mois courant. On examine également les valeurs transférées des mois précédents puisque la question du revenu n'est demandée qu'au moment où le ménage participe à l'enquête pour la première fois. Il est également impossible de choisir comme donneur des valeurs anormalement élevées ou basses, même si elles peuvent être réelles.

Les logements vacants et non existants ne contribuent pas au biais de l'échantillon. Cependant, ils produisent une hausse de la variance de l'échantillon puisqu'ils réduisent le nombre de logements dans l'échantillon de l'EPA. Une erreur peut également se produire si des logements sont

Tous les mois, les taux de glissement sont analysés en détail. Ils sont produits périodiquement au niveau des régions métropolitaines de recensement, des régions économiques et à l'échelle nationale et provinciale, pour douze groupes d'âge-sexe au Canada (15-19, 20-24, 25-29, 30-39, 40-54, 55+). Dans le cadre du dernier remaniement de l'EPA, on a récemment produit des séries révisées d'estimations commençant en 1976, qui utilisent le dénombrement du recensement de 1991, l'ajustement pour le sous-dénombrement net du recensement, un univers de population élargi (les résidents non permanents sont désormais inclus) et une nouvelle géographie. Avant le recensement de 1991, les estimations de population ne reflétaient pas la sous-couverture du recensement. Les séries du taux de glissement ont donc été révisées pour refléter ces changements, et c'est pourquoi on peut remarquer que les taux de glissement sont plus élevés qu'auparavant. Le tableau suivant contient les taux de glissement moyens pour l'année civile 1997.

Taux de glissement moyen (%) - Canada par groupes d'âge et provinces - 1997

Province		Moyenne
Canada	tous	9,3
	âges 15-19	6,1
	âges 20-24	15,6
	âges 25-29	16,1
	âges 30-39	9,8
	âges 40-54	8,0
	âges 55 +	6,9
Terre-Neuve		9,8
Île-du-Prince-Édouard		11,6
Nouvelle-Écosse		8,6
Nouveau-Brunswick		10,4
Québec		8,0
Ontario		9,7
Manitoba		6,1
Saskatchewan		10,7
Alberta		7,4
Colombie-Britannique		12,4

Tous les mois, le nombre de dossiers temporaires créés par les bureaux régionaux est contrôlé et des explications doivent être fournies en présence de toute baisse importante. Pour sélectionner son échantillon, l'EPA détient une base de données regroupant la liste des logements susceptibles d'être choisis. Six semaines avant d'entrer dans l'échantillon, la liste des logements à échantillonner est envoyée à un intervieweur sur le terrain afin que celui-ci vérifie et mette à jour la composition de celle-ci. L'échantillon est choisi à partir de cette liste au bureau central. Quoique mise à jour récemment, il est possible qu'au moment de visiter ces logements, des logements viennent s'ajouter à cette liste principalement en présence de secteurs en développement. On parle alors de dossiers temporaires, parce que pour le mois courant un numéro de dossier temporaire leur est attribué puisqu'ils ne figuraient pas dans la base de données ; la situation est rétablie le mois suivant.

Comme dernier indicateur de qualité relatif à la couverture, on produit à l'occasion la taille moyenne des ménages pour la population cible de l'EPA par province et type de secteur : rural ou urbain. On évalue si la situation présente des fluctuations ou de la stabilité.

Tous ces indicateurs permettent de déceler des problèmes potentiels relativement à la couverture de l'échantillon et de réagir en conséquence. Pour y remédier ou ralentir sa progression, on peut par exemple songer à créer des exercices à l'intention des intervieweurs pour accroître leurs connaissances des règles de composition du ménage, à distribuer un bulletin qui explique ce qu'est le glissement ou le concept de logement multiple ou encore à établir un programme de relistage d'un certain nombre de grappes jugées en pleine expansion. Le glissement sera toujours surveillé très minutieusement puisqu'en termes de qualité, il se traduit à l'EPA par une diminution de la taille de l'échantillon par rapport à ce qu'elle était lors de la mise sur pied du plan de sondage, ce qui peut introduire une source de biais possible dans les estimations. De plus, en dépit de l'application d'une méthode d'estimation pour corriger le glissement, on peut s'attendre à ce que subsiste un certain biais dans l'estimation, autre que le biais habituel d'estimation, puisque les caractéristiques des personnes et des logements omis peuvent être différentes de celles des personnes qui sont incluses dans l'échantillon.

Non-réponse

Chaque mois, durant la semaine d'enquête, les intervieweurs s'affairent à déterminer quels sont les logements sélectionnés qui contiennent des personnes

d'observations est élevé ou s'il s'agit de grands secteurs. Par contre, son effet peut être élevé lorsqu'il s'agit de petits secteurs ou bien lorsque les caractéristiques à l'étude sont rares ou rattachées à des questions délicates. Le biais non dû à l'échantillonnage, pour sa part, a tendance à se produire dans un sens plus que dans l'autre. Il peut être attribuable à la formation ou à l'attitude de l'intervieweur, à une mauvaise conception du questionnaire ou à la méthode d'imputation utilisée pour pallier la non-réponse. L'ensemble de ces facteurs peut contribuer à provoquer l'accumulation des erreurs dans une direction plutôt que dans l'autre.

La variance et/ou le biais non dus à l'échantillonnage peuvent provenir de différentes sources. Dans ce qui suit, on s'intéresse tout d'abord à la couverture, à la non-réponse, à la vacance, à la réponse et au traitement. On s'intéresse également à de nouveaux types d'indicateurs qui sont disponibles depuis l'introduction du mode d'interview assistée par ordinateur. Avec ces nouvelles mesures, il est maintenant possible de connaître certains paramètres directement liés au travail qu'effectuent les intervieweurs sur le terrain et de contrôler la performance de la nouvelle technologie qui a été adoptée.

Erreur de couverture

Les erreurs de couverture se produisent lorsque les unités d'échantillonnage de la base de sondage ne représentent pas convenablement la population cible au moment de l'enquête. Il se peut que des unités aient été omises de la base de sondage (sous-dénombrement), que des unités ne se trouvant pas dans la population cible y aient été incluses (surdénombrement) ou que des unités s'y trouvent plus d'une fois (répétitions). Toutefois, le sous-dénombrement représente le problème de couverture le plus commun. Le surdénombrement n'est pas un problème sérieux dans le cadre de l'EPA.

Les erreurs de couverture peuvent se produire à plusieurs étapes de l'enquête : pendant la conception de la base de sondage, les définitions des unités d'échantillonnage, l'attribution des probabilités de sélection aux fins de l'échantillonnage ou encore la collecte et le traitement des données. À l'EPA, l'indicateur utilisé pour mesurer l'erreur de couverture s'appelle le *taux de glissement*. Par définition, ce taux représente le pourcentage d'écart entre les estimations démographiques de l'EPA (sans données externes, c'est-à-dire basées sur les sous-poids de l'enquête) et les plus récentes estimations démographiques du recensement. Les estimations démographiques utilisées dans la détermination du taux de glissement peuvent également comporter des erreurs, et ces erreurs sont en fait un des facteurs qui contribuent au glissement. Dans le cadre de l'EPA, on observe du sous-dénombrement ; il se traduit par un taux de glissement positif. Pour réduire au maximum le biais qui en résulte, au cours du processus d'estimation, on rajuste les estimations d'échantillon en fonction de totaux de contrôle provenant de sources indépendantes.

L'omission de logements ou de personnes de la population cible, c'est-à-dire la présence de sous-couverture à l'EPA, peut introduire des erreurs non dues à l'échantillonnage. Par logement, on entend toute construction habitable répondant à certains critères. Les personnes qui vivent dans un logement composent le ménage. Il se peut qu'un logement occupé ne soit pas inscrit dans la liste des grappes pour diverses raisons : l'omission lors de l'établissement de la liste, immeuble en construction durant la dernière vérification, erreurs dans les délimitations de la grappe ou encore classifié vacant par erreur. Il est également possible que des personnes soient oubliées à l'intérieur d'un ménage, soit parce que le répondant ne révèle pas leur présence ou encore qu'on leur a attribué un lieu de résidence habituel ailleurs que dans le ménage échantillonné. Les étudiants sont souvent oubliés puisqu'ils résident ailleurs durant leurs études, quoique que leur résidence habituelle soit dans l'échantillon. Des erreurs peuvent donc se glisser dans les estimations de l'enquête, si les caractéristiques des individus non inclus dans l'enquête diffèrent de celles des individus inclus. Par exemple, si l'enquête ne rejoint pas une partie de la population qui est jeune et grandement mobile, qui affiche des taux de chômage plus élevés que la population du même âge dans l'enquête, alors le glissement biaise les estimations du chômage à la baisse. Finalement, comme mentionné précédemment, les estimations de population ont également un rôle à jouer en ce qui concerne le glissement.

D'autres facteurs pouvant contribuer au glissement de l'EPA ont été identifiés. Par exemple, la population s'accroît entre les remaniements, généralement dans des endroits spécifiques et non pas de manière uniforme. L'échantillon peut surestimer ou sous-estimer cette croissance ou en rendre compte de façon précise. Autre exemple, l'ajustement pour pallier la non-réponse (voir chapitre 5) peut également influencer le glissement. En effet, si les ménages non-répondants comptent moins de membres et s'ils sont représentés dans l'échantillon par des ménages de grande taille, alors il peut y avoir un effet sur le taux de glissement.

variation des moyennes annuelles à l'échelle nationale, provinciale et infra-provinciale. En raison des contraintes d'espace dans les publications courantes et spéciales, il n'est pas possible d'inclure les CV de toutes les estimations d'enquête publiées. Toutefois, il existe des tables de recherche qui présentent les CV approximatifs pour différents groupes d'estimations. Le tableau précédent présente quelques valeurs représentatives des coefficients de variation relatifs aux caractéristiques *employés* et *chômeurs* à l'échelle provinciale et nationale, selon les données d'enquête de janvier à juillet 1997.

On a de plus en plus porté l'attention sur la qualité des estimations relatives aux variations d'un mois à l'autre. À cet égard, le communiqué mensuel de l'EPA indique maintenant les écarts types (ET) relatifs aux variations à l'échelle provinciale et nationale pour les *employés* et les *chômeurs*. Ces chiffres sont donnés pour la période de 1997 dans le tableau ci-dessous.

Écart type de la variation d'un mois à l'autre, *employés* et *chômeurs*

Province	**ET(employés)** (milliers)	**ET(chômeurs)** (milliers)
Terre-Neuve	3	2
Île-du-Prince-Édouard	1	1
Nouvelle-Écosse	4	3
Nouveau-Brunswick	3	2
Québec	18	14
Ontario	20	15
Manitoba	4	3
Saskatchewan	3	2
Alberta	9	6
Colombie-Britannique	12	9
Canada	32	24

L'effet du plan est une autre mesure de la qualité obtenue à partir de l'échantillon. On définit cet effet comme le rapport entre la variance d'une estimation provenant d'une enquête par sondage conçue conformément à un plan donné et la variance d'une estimation qui aurait résulté d'un échantillon aléatoire simple de même taille. On peut utiliser l'effet du plan en tant qu'indice de la détérioration du plan d'échantillonnage avec le temps. Dans le cadre de l'EPA, on calcule deux types d'effet du plan, chacun dépendant des données utilisées pour l'établir. On détermine l'*effet du plan d'échantillonnage* à partir d'estimations sous-pondérées, c'est-à-dire sans pondération tenant compte des totaux de population. On calcule l'*effet du plan de sondage* au moyen de pondérations finales. Par conséquent, l'effet du plan d'échantillonnage dénote l'efficacité du plan d'échantillonnage, tandis que l'effet du plan de sondage donne une évaluation générale de la stratégie adoptée en combinant toutes les caractéristiques du plan de sondage (stratification, échantillonnage à plusieurs degrés, post-stratification et estimation). Plus l'effet est faible, plus efficace est le plan en ce qui concerne la variance de l'échantillonnage. L'observation de l'effet du plan permet donc de mesurer les changements qualitatifs subis par le plan avec le temps. Soulignons que les effets du plan d'échantillonnage sont généralement plus importants que les effets du plan de sondage basés sur les pondérations finales, puisqu'ils ne profitent pas du gain de précision qu'apporte la post-stratification.

Dans le cadre de l'EPA, on utilise l'effet du plan d'échantillonnage en conjonction avec d'autres renseignements pour décider des secteurs où une mise à jour est nécessaire. Le tableau suivant présente quelques valeurs représentatives des effets du plan d'échantillonnage et du plan de sondage pour la caractéristique chômage à l'échelle nationale et provinciale.

Effets du plan - chômeurs - 1997

Province	Employés		Chômeurs	
	Sondage	Échant.	Sondage	Échant.
Terre-Neuve	2,7	0,83	1,4	1,3
Île-du-Prince-Édouard	2,0	0,53	1,1	1,1
Nouvelle-Écosse	2,2	0,51	1,2	1,1
Noveau-Bruswick	2,0	0,56	1,4	1,4
Québec	2,1	0,55	1,1	1,0
Ontario	3,3	0,50	1,2	1,1
Manitoba	2,2	0,41	1,1	1,1
Saskatchewan	2,4	0,63	1,2	1,2
Alberta	4,1	0,40	1,1	1,0
Colombie-Britannique	2,1	0,50	1,2	1,1
Canada	2,8	0,51	1,2	1,1

Erreurs non dues à l'échantillonnage

Les erreurs non dues à l'échantillonnage peuvent survenir à toutes les étapes d'une enquête et sont causées en général par des erreurs humaines telles des erreurs d'inattention, de mauvaise compréhension ou d'interprétation. L'impact sur les estimations peut se manifester au niveau du biais et/ou de la variabilité des estimations. L'effet net de la variance non due à l'erreur d'échantillonnage peut être négligeable si le nombre

On recourt généralement à l'erreur quadratique moyenne relative à une ou plusieurs caractéristiques pour mesurer l'efficacité du plan de sondage et de la méthode d'estimation. On définit l'erreur quadratique moyenne comme la moyenne des carrés des écarts entre la valeur estimée de la caractéristique et sa valeur réelle dans la population. Dans la théorie de l'échantillonnage, pour les populations finies, la moyenne des estimations de tous les échantillons possibles est connue sous le nom d'espérance mathématique de l'estimation. On appelle biais de l'estimation la différence entre l'espérance mathématique et la valeur réelle. La variance d'une estimation d'échantillon correspond à la moyenne des carrés des écarts entre l'estimation et l'espérance mathématique. On appelle écart type de l'estimation la racine carrée de la variance.

Si la méthode d'estimation n'était pas biaisée, l'espérance mathématique de l'estimation serait identique à la valeur réelle de la caractéristique dans la population, tout comme l'erreur quadratique moyenne et la variance. Bien que certains méthodes d'estimation (comme celle qui est utilisée pour l'EPA) causent un léger biais, elles entraînent des erreurs quadratiques moyennes plus faibles que d'autres méthodes non biaisées.

L'une des principales caractéristiques d'un échantillon probabiliste, comme celui qui est utilisé dans le cadre de l'EPA, est que la variance d'un estimateur (et par conséquent son écart type) peut être estimée au moyen de l'échantillon en tant que tel. Le chapitre 5 décrit la méthode suivie pour ce faire. On utilise ici une notation simplifiée des valeurs qui y sont décrites.

Le coefficient de variation (CV) est une autre importante mesure de la qualité relativement à l'erreur d'échantillonnage. Le coefficient de variation, que l'on obtient en calculant le rapport (exprimé en pourcentage) entre écart type estimé d'une estimation et sa valeur estimée, donne le degré de fiabilité de l'estimation. Si l'on pose que Y correspond à l'estimation de la caractéristique d'intérêt et que d est l'écart type estimé de cette estimation, alors le CV est exprimé de la manière suivante : (d/Y) x 100.

L'écart type estimé (d) peut également servir à calculer l'intervalle de confiance associée à une estimation (Y). L'intervalle de confiance, qui sert à mesurer la précision, est une fonction des données de l'échantillon contenant la valeur réelle d'une caractéristique de la population observée selon un niveau de confiance donné. Si on répétait plusieurs fois l'échantillonnage, on pourrait affirmer que 95 fois sur 100, l'intervalle $Y \pm 2d$ contiendrait la valeur réelle. Dans les mêmes conditions, on pourrait affirmer que 68 fois sur 100, l'intervalle $Y \pm d$ contiendrait la valeur réelle.

Pour mettre en lumière les liens entre les différentes mesures de précision, prenons l'exemple suivant. En mars 1995, le taux de chômage de la population canadienne âgée de 15 ans et plus était de 10,8 % et l'estimation de l'écart type correspondant était de 0,0016. Par conséquent, une estimation du coefficient de variation est (0,0016/0,108) = 1,48 %. L'intervalle de confiance de 95 %, qui est calculé à partir de l'échantillon, se situe entre 10,48 % et 11,12 %, soit 0,108±0,0032. Cela signifie que, pour un niveau de confiance de 95 %, on peut affirmer que le taux de chômage de la population cible se situe entre 10,48 et 11,12 %.

Comme l'EPA est soumise à un échéancier de publication mensuel très serré, pour tout mois donné, les CV du mois courant ne sont pas disponibles pour publication immédiate. Compte tenu de la stabilité des CV observée dans le cadre de l'EPA, on fournit à la place une estimation du coefficient de variation correspondant à la moyenne des CV du semestre précédent. Ces estimations sont mises à jour deux fois par année (janvier à juin et juillet à décembre), et des facteurs de correction sont appliqués pour rendre compte de tout changement survenu (par exemple, une réduction de la taille de l'échantillon).

Grâce aux données recueillies dans le cadre de l'EPA, il est possible de produire des milliers d'estimations relatives aux caractéristiques de la population. Il s'agit d'estimations mensuelles, d'estimations de variation d'un mois à l'autre, d'estimations de moyenne de niveaux et de

Coefficients de variation mensuels observés

Province	CV (%) Employés	CV (%) Chômeurs
Terre-Neuve	2,2	6,1
Île-du-Prince-Édouard	1,7	6,5
Nouvelle-Écosse	1,2	5,3
Nouveau-Brunswick	1,2	5,5
Québec	0,79	3,5
Ontario	0,54	3,0
Manitoba	0,91	6,5
Saskatchewan	1,1	7,4
Alberta	0,76	5,9
Colombie-Britannique	0,90	5,1
Canada	0,32	1,72

CHAPITRE 6 - Qualité des données

Indicateurs de qualité à 1'EPA

Les estimations de 1'EPA, comme celles produites à l'aide de tout autre enquête-échantillon, peuvent comporter des erreurs d'échantillonnage et des erreurs non dues à l'échantillonnage. Conséquemment, pour interpréter correctement les estimations de cette enquête, il faut une connaissance de leur qualité.

Dans une enquête par sondage, des inférences sont faites au sujet de la population visée à partir des données recueillies auprès d'une partie seulement (échantillon) de cette population. Les résultats sont probablement différents de ceux qu'on obtiendrait si on menait un recensement complet de cette population dans les mêmes conditions. L'erreur due au fait d'étendre à toute la population des conclusions fondées sur un échantillon seulement est appelée erreur d'échantillonnage. Au nombre des facteurs qui contribuent aux erreurs d'échantillonnage figurent : la taille de l'échantillon, la variabilité des caractéristiques étudiées, le plan d'échantillonnage et la méthode d'estimation.

L'erreur non due à l'échantillonnage, comme son nom l'indique, n'a rien a voir avec le processus d'échantillonnage et se produit dans un recensement (auquel participent toutes les unités de la population) aussi bien que dans une enquête par sondage. Ce type d'erreur peut survenir à n'importe quelle étape d'une enquête (planification, conception, collecte des données, codage, saisie, vérification, estimation, analyse et diffusion des données) et est principalement attribuable à des erreurs humaines. On peut également associer l'erreur non due à l'échantillonnage à d'autres types d'erreurs comme par exemple à des erreurs dans les sources d'information et les méthodes utilisées pour obtenir des projections de population, à des erreurs de rajustements saisonniers, etc. Pour assurer et contrôler la qualité de ses données, l'EPA s'est dotée d'un programme poussé sur la qualité des données. Toute une gamme d'indicateurs de qualité sont produits sur une base régulière et analysés avec soin. En présence de valeurs inhabituelles, les responsables des activités concernées de l'EPA sont immédiatement avisés afin de garantir la qualité des données d'une enquête à l'autre. Par ailleurs, certains indicateurs sont contrôlés d'une façon moins régulière puisque leur rôle est de permettre de déceler des tendances ou des effets à long terme, par exemple les conséquences de certains changements d'ordre opérationnel ou apportés au plan de sondage. Ces renseignements à long terme au sujet de la fiabilité des données peuvent servir à apporter des changements susceptibles d'améliorer la qualité générale des résultats et d'aider les analystes et les utilisateurs de données, tant à l'interne qu'à l'externe, dans leur travail. Dans les lignes qui suivent, les indicateurs de qualité produits pour l'EPA sont présentés sous deux rubriques : erreur d'échantillonnage et erreurs non dues à l'échantillonnage.

Erreurs d'échantillonnage

Les répercussions d'une erreur d'échantillonnage sur les estimations de l'enquête sont fonction de plusieurs facteurs. Le plus évident est la taille de l'échantillon. Tout autre facteur étant constant, l'erreur d'échantillonnage diminue généralement au fur et à mesure que la taille de l'échantillon augmente. Outre la taille d'échantillon, l'erreur d'échantillonnage dépend de facteurs tels la variabilité de la population, la méthode d'estimation et le plan de sondage. Pour un échantillon d'une taille donnée, l'erreur d'échantillonnage est liée à diverses caractéristiques du plan de sondage comme la méthode de stratification utilisée, la répartition de l'échantillon, le choix des unités d'échantillonnage et la méthode de sélection employée à chaque degré d'échantillonnage pour un plan à plusieurs degrés. De même, pour un plan de sondage donné, la méthode d'estimation utilisée joue un rôle important. Enfin, même en posant une taille d'échantillon, un plan d'échantillonnage et des méthodes d'estimation identiques, l'évaluation de différentes caractéristiques (pour lesquelles des données ont été recueillies à partir du même échantillon) produirait des erreurs d'échantillonnage différentes, puisque le degré de variabilité varierait d'une caractéristique à une autre. Ces erreurs sont généralement plus grandes pour les caractéristiques qui sont relativement rares ou qui sont distribuées de façon non égale dans l'ensemble de la population que pour les caractéristiques plus courantes et plus homogènes. Ainsi, bien qu'elles se fondent sur le même échantillon, les estimations relatives au chômage comportent généralement une erreur d'échantillonnage plus élevée que les données relatives à l'emploi.

variation mensuelle, le gain peut être beaucoup plus important. Ainsi, la variance de l'estimation de variation mensuelle de l'emploi en Ontario est réduite de moitié. Dans le cas de la variation de l'emploi dans certaines industries, la réduction de la variance est encore plus grande. Une des conséquences importantes de ce dernier résultat est que certaines séries chronologiques qui ne pouvaient pas être corrigées des fluctuations saisonnières de manière efficace dans le passé pourront l'être avec l'adoption de la méthode de l'analyse de régression modifiée, c'est-à-dire que cette méthode augmente suffisamment le rapport signal/bruit pour permettre à la méthode de désaisonnalisation de détecter la structure saisonnière. Étant donné ces résultats encourageants, il est prévu d'introduire la méthode de l'analyse de régression modifiée dans l'EPA dans un avenir proche.

Pondération pendant l'introduction de l'échantillon

On a introduit le nouveau plan de sondage de l'EPA en remplaçant l'ancien échantillon un groupe de renouvellement à la fois, sur une période de six mois. À chaque fois que la participation de ménages sélectionnés suivant l'ancien plan de sondage était échue, ils étaient remplacés par des ménages prélevés suivant le nouveau plan de sondage. Ce processus a commencé en octobre 1994, et en mars 1995 l'échantillon était complètement renouvelé. Étant donné les modifications apportées au système de numérotation de l'EPA, au découpage géographique infraprovincial et à la méthode de pondération, la pondération a dû être envisagée d'une manière spéciale.

- Aucune stabilisation du nouvel échantillon n'a été effectuée pendant la période d'introduction.

- On a décidé de cesser d'utiliser le facteur rural/urbain dès octobre 1994. Comme nous l'avons mentionné précédemment, ce facteur n'est pas nécessaire pour le nouvel échantillon. Son application à l'ancien échantillon pendant la phase d'introduction ne pouvait que produire des facteurs de pondération instables. À mesure que l'ancien échantillon était retiré, le déséquilibre entre les effectifs ruraux et urbains de l'ancien échantillon aurait été accentué, car l'apport du nouvel échantillon à ceux-ci n'aurait pas été pris en compte.

- On a utilisé l'ancienne méthode de compensation de la non-réponse pour l'ancien échantillon et la nouvelle méthode pour le nouvel échantillon.

- L'utilisation de répliques de secteurs spéciaux a été maintenue uniquement dans le cas de l'ancien échantillon.

- Une fois terminé le calcul des sous-poids, les deux échantillons ont été combinés pour l'étape de pondération finale.

Nouvelles techniques : Estimation composite

Jusqu'à maintenant, on n'avait pas encore exploité le fait que les cinq sixièmes de l'échantillon de l'EPA coïncident d'un mois d'enquête à l'autre pour améliorer les estimations . Il est bien connu que, dans un plan d'enquête avec renouvellement de l'échantillon, on peut se servir de l'échantillon commun pour produire une meilleure estimation de la variation qu'en calculant simplement la différence entre deux estimations de deux mois consécutifs. À son tour, cette amélioration de la méthode d'estimation permet d'améliorer l'estimation de niveau. Par exemple, l'estimateur composite k conventionnel est une combinaison linéaire de l'estimation de niveau courante, disons un estimateur de régression, et d'une autre estimation de niveau obtenue en prenant l'estimation de niveau du mois précédent et en la mettant à jour en prenant une estimation de la variation basée sur l'échantillon commun, à savoir

$$\text{est}'_{(t+1)} = K \times \text{est}_{(t+1)} + (1\text{-}K) \times [\text{est}'_{(t)} + \text{change}_{\text{commun}}],$$

où le nombre premier désigne une estimation composite. Malgré que les estimateurs composites conventionnels permettaient d'améliorer les estimations, ils comportaient plusieurs inconvénients, comme les problèmes de cohérence des estimations. On a donc opté jusqu'à présent de ne pas utiliser l'estimation composite pour l'EPA.

L'article de Singh et coll. (1997) décrit une version de l'estimation composite appelée estimation composite par analyse de régression modifiée. L'estimateur par analyse de régression modifiée ressemble dans sa conception aux estimateurs composites conventionnels mais diffère de ceux-ci dans le détail. En particulier, il traite en même temps toutes les caractéristiques à intégrer et offre une solution au problème de cohérence. La méthode de l'analyse de régression modifiée a l'avantage pratique de bien cadrer avec le système d'estimation utilisé présentement dans l'EPA, car les caractéristiques à l'étude entrent dans la procédure d'estimation en tant que totaux de contrôle. Elle possède également deux propriétés essentielles : chaque ménage de l'échantillon a un seul poids (c.-à-d. que le poids ne dépend pas de la caractéristique à l'étude) et les parties s'additionnent au total correspondant (p. ex., la somme des caractéristiques *personnes occupées* et *chômeurs* est encore égale à la taille de la population active, ce qui n'est pas le cas de la méthode conventionnelle où chaque variable est traitée séparément).

Quant aux caractéristiques contrôlées dans l'analyse de régression modifiée, la mesure de leur variance permet de constater une grande amélioration de l'efficacité. Ainsi, d'après les études que nous avons effectuées, dans le cas des estimations d'emploi dans certaines industries dont les estimations de régression sont instables, le gain d'efficacité dépasse parfois 40 pour 100. Dans le cas des estimations provinciales d'emploi et de chômage, les gains sont plus modestes mais restent appréciables. Par exemple, en Ontario ils sont de cinq et de douze pour cent, respectivement. Dans le cas des estimations de

l'élimination du facteur rural-urbain, le changement de la définition du secteur de non-réponse, le changement de la définition du secteur de stabilisation et l'élimination de la répétition des enregistrements pour les secteurs spéciaux. Dans ce qui suit se trouve une discussion portant sur les méthodes utilisées pour adapter la pondération pendant l'introduction du nouvel échantillon.

Facteur rural-urbain. Dans l'ancien plan de sondage de l'EPA, certaines strates de secteurs non autoreprésentatifs comprenaient à la fois des parties rurales et urbaines, ce qui pouvait entraîner une surreprésentation ou une sous-représentation de la population rurale ou urbaine. On utilisait un facteur de rajustement du sous-poids afin que la proportion de population rurale et urbaine de chaque région économique corresponde à celle du recensement de 1981. Dans le plan de sondage actuel, la stratification est explicite, ce qui fait que l'échantillon est représentatif. Ce facteur n'est donc plus nécessaire.

Compensation pour la non-réponse. La plupart des cas de non-réponse, dans l'EPA, sont réglés par un rajustement du poids. L'application d'un tel facteur de compensation repose sur l'hypothèse que les non-répondants peuvent être représentés par leurs homologues répondants dans lesdits secteurs de non-réponse. L'ancienne méthode de définition d'un secteur de non-réponse consistait à la considérer comme une strate, dans les secteurs autoreprésentatifs ou spéciaux, et comme la partie rurale ou urbaine d'une unité primaire d'échantillonnage, dans les secteurs non autoreprésentatifs.

Dans le cadre du programme de contrôle de la qualité, on a surveillé l'évolution des schémas de non-réponse à l'enquête et on a remarqué depuis longtemps que les schémas de non-réponse ne sont pas les mêmes d'un groupe de renouvellement à l'autre. La durée de la participation d'un ménage à l'enquête a une incidence sur l'ampleur de la non-réponse. Les proportions des différents types de non-répondants (aucun contact, absence temporaire et refus) varient également suivant la durée de la participation à l'enquête. On a donc décidé d'inclure la durée de la participation à l'enquête dans la définition du secteur de non-réponse. Si cet élément avait été ajouté à l'ancienne définition sans autre modification, le rendement de l'échantillon aurait été trop faible dans certains secteurs pour permettre un rajustement. La définition a été modifiée de manière à inclure tous les ménages qui appartiennent à une même région économique d'assurance-emploi et au même type de base et qui participent à l'enquête depuis le même nombre de mois.

Secteurs de stabilisation. Comme dans le cas des secteurs de non-réponse, un changement de la définition des secteurs de stabilisation a été introduit. Ces secteurs étaient auparavant définis comme étant toutes les strates ayant le même poids de base dans une province. Ils sont maintenant définis comme étant tous les ménages qui appartiennent à une même région économique d'assurance-emploi et qui participent depuis le même temps à l'enquête. On rajuste ensuite le poids dans le secteur de stabilisation, en regroupant toutes les strates qui ont la même fraction de sondage. Ce changement reflète l'importance accrue accordée aux REAE dans le remaniement du plan d'échantillonnage.

Répliques de secteurs spéciaux. L'ancien plan de sondage de l'EPA comprenait trois bases comportant un effectif de population plutôt restreint. Il s'agissait de la base institutionnelle, de celle des régions éloignées et de celle des régions éloignées du Québec. Ensemble, ces bases comprenaient environ deux pour cent de la population. Le coût des interviews y était beaucoup plus élevé que dans les autres régions, et le rendement de l'échantillon était en général très faible. Ce dernier inconvénient avait mené à l'attribution de petites tâches d'intervieweurs répartis sur de grands territoires. Vu les problèmes de gestion soulevés par l'attribution de tâches réduites et le peu d'importance numérique de la population concernée, on a décidé de ne pas tenir compte des régions infraprovinciales dans l'échantillonnage de ces secteurs. Par exemple, il n'était pas possible de prélever un échantillon représentatif de ménages de secteurs spéciaux dans toutes les régions économiques, mais chaque province pouvait en fournir un. Pour remédier à cette situation pendant l'estimation, on a reproduit tous les enregistrements des secteurs spéciaux dans l'ensemble des régions économiques qui comptaient des effectifs du même type. Ensuite, les poids provinciaux ont été proportionnés de manière à représenter la région infraprovinciale. Par exemple si une région économique contenait dix pour cent de la population éloignée de la province, les enregistrements de l'échantillon des régions éloignées de cette région ont été reproduits et leur poids de base a été multiplié par 0,1.

Dans le nouveau plan de sondage, la population des secteurs institutionnels n'est pas échantillonnée à partir d'une base spéciale. Cet effectif n'est plus considéré comme un cas spécial. Le reste de la base des secteurs spéciaux est échantillonné de la même façon que dans l'ancien plan. Étant donné le faible impact sur les estimations, la reproduction des enregistrements n'est plus nécessaire.

$$J = \sum_{h=1}^{H} J_h$$

où **H** est le nombre total de strates de l'échantillon.

(ii) Dans la strate donnée, on rajuste les sous-poids de tous les ménages des répliques J_h **- 1** restantes afin de compenser l'abandon des ménages de l'UPD retirée. Le poids rajusté est

$$a_k^{adj} = \frac{J_h}{(J_h - 1)} a_k$$

(iii) Pour le reste de l'échantillon dont les sous-poids ont été rajustés, on recalcule les poids finaux de manière à obtenir une nouvelle estimation de la caractéristique désirée. La nouvelle estimation peut être indiquée comme suit :

$$\hat{t}_{yr(ha)}$$

La notation **(ha)** indique que la **a**ième réplique de la **h**ième strate a été retirée afin d'obtenir la nouvelle estimation. Par conséquent, l'estimation précédente est basée sur toutes les répliques sauf la **(ha)**ième.

Cette opération est répétée pour chaque réplique de l'échantillon. On obtient alors **J** différentes estimations de la caractéristique désirée. La variance de l'estimation d'un total est fournie par la formule :

$$\hat{V}(\hat{t}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{t}_{yr(ha)} - \hat{t}_{yr})^2$$

La question tourne souvent autour du rapport entre deux totaux. Par exemple, le taux de chômage est le rapport du nombre total de chômeurs à la population active totale exprimé en pourcentage. En général, on utilise pour un rapport de **100(y/z)%** la formule de variance :

$$\hat{V}\left(100\frac{\hat{t}_{yr}}{\hat{t}_{zr}}\right) = (100)^2 \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} \left(\frac{\hat{t}_{yr(ha)}}{\hat{t}_{zr(ha)}} - \frac{\hat{t}_{yr}}{\hat{t}_{zr}}\right)^2$$

Les variances pour les estimations de changements mensuels et des moyennes sur plusieurs mois nécessitent l'établissement d'un lien temporel entre les estimations par la méthode du jackknife. Prenons l'estimation par la différence

$$\hat{D}_{yr} = \hat{t}_{yr}^{2} - \hat{t}_{yr}^{1}$$

et les estimations par la méthode du jackknife correspondantes

$$\hat{D}_{yr(ha)} = \hat{t}_{yr(ha)}^{2} - \hat{t}_{yr(ha)}^{1}$$

où les exposants désignent des mois consécutifs. L'estimation de la variance est fournie par la formule :

$$\hat{V}(\hat{D}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{D}_{yr(ha)} - \hat{D}_{yr})^2$$

La variance des moyennes est obtenue de la même manière. Soit la moyenne sur **n** mois

$$\hat{A}_{yr} = \sum_{i=1}^{n} \frac{\hat{t}_{yr}^{i}}{n}$$

et les estimations par la méthode du jackknife

$$\hat{A}_{yr(ha)} = \sum_{i=1}^{n} \frac{\hat{t}_{yr(ha)}^{i}}{n}$$

L'estimation de la variance est fournie par la formule :

$$\hat{V}(\hat{A}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{A}_{yr(ha)} - \hat{A}_{yr})^2$$

Modifications apportées à la méthodologie précédente

À l'occasion du remaniement du plan de sondage, plusieurs changements ont été apportés à la méthode de pondération de l'EPA. Les plus importants ont été

Le facteur $\mathbf{B}_q$ sera défini plus loin. D'après la formule ci-dessus, nous voyons que l'estimateur de régression peut être considéré comme un estimateur avec sous-pondération auquel on a ajouté un facteur de compensation. Si l'estimation basée sur l'échantillon s'approche du total connu pour $\mathbf{x}_q$, alors le facteur de compensation tend vers zéro. Si ces deux valeurs sont différentes, le facteur de compensation devient important. L'utilisation de l'estimateur de régression permet d'obtenir de meilleurs résultats si les caractéristiques sont en corrélation avec les variables auxiliaires.

Pour définir le facteur $\mathbf{B}_j$, on emploie la notation matricielle :

$$\hat{\mathbf{B}} = (\hat{B}_1,\ldots,\hat{B}_Q)' = \Big(\sum_{k=1}^{n} \frac{x_k x_k^t a_k}{c_k}\Big)^{-1} \sum_{k=1}^{n} \frac{x_k y_k a_k}{c_k}$$

où

$$\Big(\sum_{k=1}^{n} \frac{x_k x_k^t a_k}{c_k}\Big)^{-1}$$

est une matrice $\mathbf{Q}\times\mathbf{Q}$. Il s'agit de la matrice inverse de la somme pondérée des produits croisés de l'estimation par régression, et

$$\sum_{k=1}^{n} \frac{x_k y_k a_k}{c_k}$$

est un vecteur $\mathbf{Q}\times\mathbf{1}$.

L'estimateur ci-dessus peut être modifié comme suit :

$$\hat{t}_{yr} = \sum_{k\in s} y_k a_k g_k$$

où

$$g_k = 1 + (t_x - \hat{t}_{xa})^t \Big(\sum_{k\in s} \frac{x_k x_k^t a_k}{c_k}\Big)^{-1} \frac{x_k}{c_k}$$

Les facteurs g_k ou *facteurs g* sont des facteurs appliqués aux sous-poids afin d'obtenir les poids finaux.

Le fait que les poids finaux ne dépendent pas de la caractéristique y signifie que le même poids peut être utilisé pour le calcul de toutes les caractéristiques à l'étude. Soulignons également que le facteur g est défini au niveau ménage et que le même facteur est attribué à chaque membre d'un ménage. Les facteurs g pourraient aussi bien être calculés par la formule :

$$g_i = 1 + (t_x - \hat{t}_{xa})^t \Big(\sum_{i\in s} z_i z_i^t a_i\Big)^{-1} z_i$$

où a_i est le sous-poids attribué à la $\mathbf{i}^{\text{ième}}$ personne de l'échantillon et où l'indicateur $\mathbf{z}_i$ contient, pour chaque personne, la moyenne des valeurs des indicateurs pour chaque personne d'un même ménage, à savoir :

$$z_i = \frac{1}{c_k} \sum_{i}^{c_k} y_i$$

La même valeur de l'indicateur **z**, à savoir le moyenne du ménage, est attribuée à chaque personne **i** du ménage **k**.

Estimation de la variance: l'algorithme du jackknife

L'estimateur de variance utilisé dans l'EPA est le jackknife. Une description de son application générale est donnée dans Wolter (1985). Nous traitons ici de son application dans l'EPA. La première étape de la méthode du jackknife consiste à créer des répliques d'échantillon à partir des données de l'EPA. Dans chaque strate du plan de sondage, on prélève à tour de rôle une unité d'échantillonnage du premier degré. Cette UPD est retirée de l'échantillon et les sous-poids du reste de la strate sont rajustés pour compenser ce retrait. On recalcule ensuite les estimations finales en utilisant les répliques, c'est-à-dire l'échantillon provincial moins les UPD retirées. En répétant cette opération pour chaque UPD de l'échantillon, on obtient des estimations pour toutes les répliques, c'est-à-dire autant d'estimations qu'il y a d'UPD. La variabilité entre les estimations de réplique sert à évaluer la variance de l'estimation d'échantillon. Pour simplifier, convenons d'appeler «réplique» chaque UPD retirée.

Pour obtenir une variance par la méthode du jackknife, on procède comme suit :

(i) Retirer tous les ménages d'une réplique déterminée. Soit les répliques $\mathbf{a} = 1, \ldots, \mathbf{J}_h$. C'est-à-dire que la $\mathbf{h}^{\text{ième}}$ strate contient $\mathbf{J}_h$ répliques, toutes désignées par **a**.

Le nombre total de répliques de l'échantillon est

Le sous-poids est défini comme étant le produit du poids de sondage et du facteur de compensation de la non-réponse :

$$a_k = f_{puf..r} \times \pi_k^{-1}$$

Il est à noter que le même sous-poids est attribué à tous les membres d'un même ménage.

Comme nous l'avons déjà mentionné, nous pouvons nous servir du sous-poids pour évaluer les caractéristiques voulues. Étant donné la caractéristique Y, disons *l'emploi*, nous voulons obtenir le nombre total de personnes occupées dans la population. Celle-ci peut être exprimée par :

$$t_y = \sum_U y_i$$

où la sommation de U correspond à une sommation de toutes les personnes de la population admissible (l'indice **i** ci-dessus désigne des particuliers) et où y_i prend la valeur un si un individu **i** est employé et la valeur zéro dans le cas contraire.

L'estimation sous-pondérée définie ci-dessus serait la suivante :

$$\hat{t}_{ya} = \sum_s y_i a_i$$

où la sommation de **s** correspond uniquement à celle des personnes de l'échantillon et où a_i est le sous-poids. Il est utile de mentionner que, dans certains cas, la formule ci-dessus peut être réécrite comme suit :

$$t_y = \sum_{k=1}^{N} \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{N} y_k$$

et

$$\hat{t}_{ya} = \sum_{k=1}^{n} a_k \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{n} y_k \, a_k$$

où $\mathbf{c_k}$ est le nombre de personnes dans le ménage **k**, **N**, le nombre de ménages dans la population et **n**, le nombre de ménages dans l'échantillon. Les valeurs $\mathbf{y_k}$ représentent les nombres totaux des ménages $\sum_{i \in k} y_i$ qui possèdent les caractéristiques à l'étude, comme le nombre de personnes ayant un emploi dans le ménage. Il importe de se rappeler que l'indice k indique la somme des valeurs d'un ménage et l'indice i, la valeur d'un particulier, et comme abus de notation i a été utilisé plutôt que ki.

L'EPA dispose d'estimations de population postcensitaires calculées indépendamment de l'échantillon. Celles-ci sont utilisées comme données auxiliaires dans le calcul d'un ensemble de poids finals. Afin de tirer profit de ces données auxiliaires, des méthodes telles la post-stratification ou un estimateur de régression peuvent être utilisées. L'approche de régression utilisée dans l'EPA est décrite dans l'article de Lemaître et Dufour (1987). La méthode décrite ci-après est tirée de Särndal et coll. (1992).

Pour commencer, prenons la notation suivante :

- y_i est la valeur de la caractéristique à l'étude se rapportant au particulier i.
- y_k est le total des valeurs pour un ménage de la caractéristique à l'étude se rapportant au ménage k.
- Q est le nombre des variables auxiliaires utilisées dans l'estimation. Chacune des variables auxiliaires est indiquée par q = 1,...,Q.
- x_{qi} est la valeur du $q^{ième}$ indicateur de l'individu i. L'indicateur prend la valeur un si l'individu i appartient à la $j^{ième}$ catégorie auxiliaire et la valeur zéro dans le cas contraire.
- x_{qk} est le total des valeurs du $q^{ième}$ indicateur de tous les membres du ménage k.
- x_k est un vecteur Q x 1 dont la $q^{ième}$ composante est le total x_{qk} du ménage correspondant.
- c_k est la taille du $k^{ième}$ ménage.
- $\hat{t}_{ya}$ est l'estimation sous-pondérée décrite ci-dessus.
- t_{xq} est le chiffre de population connu pour la $q^{ième}$ variable auxiliaire.
- $\hat{t}_{x_q a}$ est l'estimation sous-pondérée pour la $q^{ième}$ variable auxiliaire.

Par conséquent

$$\hat{t}_{x_q a} = \sum_s x_{qi} \, a_i$$

L'estimateur de régression peut être formulé comme suit :

$$\hat{t}_{yr} = \hat{t}_{ya} + \sum_{q=1}^{Q} \hat{B}_q \, (t_{x_q} - \hat{t}_{x_q a})$$

$$c_{pufh'} = \frac{R^{*}_{pufh'}}{R_{pufh}}$$

Il est également nécessaire de rajuster le reste des grappes de l'ancienne strate pour compenser la perte d'une grappe. Rappelons que la fraction de sondage de l'ancienne strate est $\mathbf{R_{pufh}}$. Soit $\mathbf{R^{R}_{pufh}}$ la fraction de sondage qui doit être appliquée au reste de l'ancienne strate pour obtenir le rendement prévu des grappes restantes. On obtient le facteur suivant, qui est appliqué à tous les ménages du reste de la strate :

$$c_{pufh} = \frac{R^{R}_{pufh}}{R_{pufh}}$$

Méthode III: Sous-échantillonnage de grappe

Il s'agit du cas le plus simple. Les ménages sélectionnés sont sous-échantillonnés et seuls les ménages sous-échantillonnés sont interviewés. Si $\mathbf{R_{pufh.j}}$ est la fraction de sondage initiale de la grappe et $\mathbf{R^{*}_{pufh.j}}$, la fraction de sondage de grappe qui permet d'obtenir le degré de sous-échantillonnage approprié, le sous-poids de grappe est :

$$c_{pufh.j} = \frac{R^{*}_{pufh.j}}{R_{pufh.j}}$$

Comme nous l'avons mentionné précédemment, les poids de stabilisation sont calculés dans les secteurs de stabilisation. Dans le présent plan de sondage, un secteur de stabilisation est défini comme étant l'ensemble des strates qui appartiennent à la même REAE. On divise ensuite ce secteur en groupes de renouvellement communs. Dans chaque secteur de stabilisation, on détermine une taille d'échantillon de base. Il s'agit du nombre de ménages qui devraient être prélevés dans le secteur en fonction de la répartition de l'échantillon. Ce nombre est représenté par la valeur $\mathbf{b_{pu..r}}$. Lorsque l'échantillonnage est effectué, on obtient effectivement un nombre donné de ménages, disons $\mathbf{n_{pu..r}}$. Si $\mathbf{n_{pu..r} > b_{pu..r}}$, c'est que le secteur est suréchantillonné, et l'excédent de ménages est abandonné aléatoirement, par un échantillonnage systématique. Comme les grappes qui ont été sous-échantillonnées par l'application de la méthode III susmentionnée ne peuvent pas être soumises à une stabilisation, elles sont exclues du calcul du poids de stabilisation. Soit $\mathbf{c_{pu..r}}$ le nombre total de ces logements dans le secteur de stabilisation.

Lorsqu'on procède à la stabilisation d'un secteur, le facteur g suivant est appliqué aux ménages de ce secteur :

$$s_{pu..r} = \frac{n_{pu..r} - c_{pu..r}}{b_{pu..r} - c_{pu..r}}$$

Il est à noter qu'aucun poids de stabilisation n'a été attribué à certains ménages du secteur de stabilisation. Ces ménages ont été définis précédemment. Il s'agit essentiellement de ménages qui ne pouvaient pas être retirés dans un processus de stabilisation.

Le poids de sondage de chaque ménage se calcule alors par la formule :

$$\pi^{-1}_{pufhrjk} = w_{pufh} \times c_{pufh.j} \times s_{pu..r}$$

Le poids de sondage est l'inverse de la probabilité d'inclusion d'un ménage donné. Dans la notation subséquente du poids de sondage, on adoptera un indiçage plus simple, à savoir :

$$\pi^{-1}_{k} = \pi^{-1}_{pufhrjk}$$

Le prochain rajustement à effectuer est celui de la non-réponse. On définit les secteurs de non-réponse et on applique un facteur de compensation pour tenir compte de la non-réponse des ménages. Dans l'EPA, les secteurs de non-réponse sont définis comme étant tous les ménages échantillonnés qui appartiennent à la même REAE, au même type de base et au même groupe de renouvellement. Le facteur de compensation est égal au rapport pondéré du nombre de ménages échantillonnées au nombre de ménages répondants, à savoir :

$$f_{puf.r} = \frac{\sum_{k \in s} \pi^{-1}_{k}}{\sum_{k \in r} \pi^{-1}_{k}}$$

où la sommation de s correspond à la sommation de tous les ménages du secteur de non-réponse et où la sommation de r, à celle de tous les ménages répondant du secteur. Le même facteur de compensation de la non-réponse est appliqué à tous les ménages d'un même secteur de non-réponse.

Dans l'EPA, le poids est attribué aux ménages. L'ERG permet de calculer, pour chaque ménage, un poids final qui fait concorder la somme des poids finaux des membres de l'échantillon appartenant à un groupe d'âge-sexe particulier ou à une région infraprovinciale déterminée avec les estimations de population utilisées comme données auxiliaires. En outre, les estimations des *personnes occupées*, des *chômeurs* et des *inactifs* sont rapportées aux chiffres de population utilisés comme données auxiliaires, car tous les membres de l'échantillon font partie d'une de ces trois catégories. Comme tous les membres d'un même ménage ont le même poids, les estimations relatives aux familles concordent avec les estimations relatives aux personnes. La méthodologie utilisée avant l'adoption de l'estimateur de régression de Lemaître-Dufour ne permettait pas de faire cela.

Pour conclure, mentionnons certains avantages offerts par l'étape de calcul du poids final :

- concordance des estimations avec les estimations démographiques,
- rajustement en fonction de l'erreur de couverture,
- attribution du même poids à tous les membres d'un ménage,
- réduction de l'erreur d'échantillonnage liée aux estimations.

Notation algébrique de la pondération d'un enregistrement

Voici une notation algébrique de la pondération. Commençons par expliquer la notation. Le plan de sondage de l'EPA correspond à une hiérarchie de zones géographiques emboîtées.

Soit p=1, ... , 10 la province;
u=1, ... , U la REAE u dans la province p;
f=1, ... , F le type de base de la REAE u;
h=1, ... , H la strate h dans la base f;
r=1, ... , 6 le groupe de renouvellement de la strate h;
j=1, ... , J la grappe j du groupe r;
k=1, ... , K le ménage k de la grappe j et
i=1, ... , c_k le membre i du ménage k.

Avec cette notation, un ménage est indiqué par l'indice **pufhrjk** . Un indice qui contient des points ou qui ne contient pas certains indices indique un renvoi à un niveau de totalisation. Par exemple, l'indice **pu..r** désigne tous les ménages du groupe de renouvellement r de la REAE u de la province p, en regroupant les ménages aux niveaux situés au-dessus des indices manquants.

Parfois, les strates recoupent les limites des REAE. La plupart du temps, c'est parce que DRHC a redéfini ses REAE après le remaniement de l'EPA. Des techniques spéciales d'estimation sont utilisées pour produire les estimation relatives à ces régions. Notons que ces zones géographiques ne sont pas parfaitement emboîtées. Cela ne pose aucun problème en ce qui concerne les méthodes d'estimation standard décrites à la présente section.

Lorsque le plan de sondage est établi, les probabilités de sélection inverses de tous les ménages d'une même strate sont identiques. Le poids de base peut être indiqué comme suit :

$$\mathbf{w}_{pufh}$$

Les deux facteurs de pondération suivants, le sous-poids de grappe et le poids de stabilisation, rajustent le poids de base pour tenir compte des diverses corrections à apporter aux rendements de l'échantillon décrites plus haut. La méthode de calcul du sous-poids de grappe dépend de la méthode de sous-échantillonnage utilisée.

Méthode I : Sous-échantillonnage de secteur

Dans ce cas, la grappe est subdivisée en plus petites grappes. Un échantillon de grappes est ensuite prélevé, puis échantillonné afin d'obtenir une forme de rendement global fixe. Si la fraction de sondage de l'ancien échantillon était $\mathbf{R}_{pufh.j}$ et que la fraction de sondage qui a dû être appliquée à l'échantillon initial pour obtenir le nouveau rendement global est $\mathbf{R}^{*}_{pufh.j}$, le sous-poids de grappe est :

$$c_{pufh.j} = \frac{R^{*}_{pufh.j}}{R_{pufh.j}}$$

Méthode II: Grappe autoreprésentative

Dans ce cas, la grappe où la population s'est accrue est retirée de la strate et forme une nouvelle strate. La nouvelle strate **h'**, disons, est subdivisée en grappes et échantillonnée. Si on appliquait aux nouvelles strates la même fraction de sondage qu'à l'ancienne strate, il ne serait pas nécessaire d'appliquer un facteur de compensation. Par contre, le rendement serait probablement très faible. Si la fraction de sondage de l'ancienne strate était $\mathbf{R}_{pufh}$ et que la fraction de sondage de la nouvelle strate est $\mathbf{R}^{*}_{pufh'}$, le sous-poids de grappe applicable aux ménages de la nouvelle grappe n'est que :

répondants possèdent des caractéristiques semblables à celles des non-répondants. Dans l'EPA, le secteur de non-réponse est défini comme étant tous les ménages qui appartiennent à la même REAE et au même type de secteur et qui font partie de l'échantillon depuis le même nombre de mois. Par type de secteur, on entend le type base de prélèvement de l'échantillon (voir le chapitre 2). Leur classification est la suivante :

RMR avec base d'appartements,
RMR normale,
autre que RMR – PPCAO,
SD urbain,
grappes urbaines,
plan à trois degrés urbain,
SD rural,
plan à trois degrés rural,
région éloignée.

La durée de participation à l'enquête est incluse dans la définition du secteur de non-réponse parce qu'on sait que l'ampleur et les schémas de non-réponse (refus, absence de contact, etc.) varient suivant la durée de la participation d'un ménage à l'enquête. La relation de cette question avec la compensation de la non-réponse est traitée dans Kennedy et coll. (1994) . Une des innovations du nouveau plan de sondage est la formation de strates de ménages à revenu élevé. À cause de leurs caractéristiques uniques, celles-ci sont traitées comme des secteurs de non-réponse à part. Il est à noter que les secteurs de non-réponse ne se chevauchent pas et que, réunis, ils couvrent l'ensemble de la population-cible.

Dans chaque secteur de non-réponse, un *facteur de compensation de la non-réponse* est calculé. Le facteur de compensation d'un secteur de non-réponse est défini comme étant le rapport du nombre de ménages échantillonnés, pondérés par l'application du poids de sondage afin qu'ils représentent les ménages du secteur, au nombre de ménages répondants pondérés de manière à obtenir une estimation du nombre de ménages du secteur qui devraient répondre. Si **n** est le nombre de ménages échantillonnés du secteur de non-réponse **b** et **r,** le nombre de ménages répondants, le facteur de compensation de la non-réponse fourni par la formule :

$$f_b = \frac{\sum_{k=1}^{n} \pi_k^{-1}}{\sum_{k=1}^{r} \pi_k^{-1}}$$

où π_k^{-1} est le poids de sondage attribué au ménage.

Comme il n'est pas souhaitable que le poids ainsi obtenu ait une valeur supérieure à deux, le secteur de non-réponse est rattaché, le cas échéant, à un autre secteur de non-réponse choisi de manière à ce que le poids commun résultant soit inférieur à deux. Le secteur de non-réponse qui doit faire l'objet d'un regroupement doit provenir de la même province, du même type de base et du même groupe de renouvellement (en effectuant un regroupement entre REAE au besoin).

Ce facteur de pondération est appliqué à tous les ménages répondants du secteur. Le *sous-poids* est défini comme étant le produit du poids de sondage **w** et du facteur d'ajustement de non-réponse $\mathbf{f_b}$.

Poids final

En théorie, le sous-poids défini ci-dessus pourrait servir à produire les estimations des caractéristiques désirées. Cependant, on sait, d'après la théorie de l'estimation, que s'il existe des données auxiliaires sur la population-cible et que celles-ci sont en corrélation avec les caractéristiques recherchées, elles peuvent servir à obtenir des estimations plus justes. Prenons un échantillon qui, par hasard, serait composé de 50 % de femmes et de 50 % d'hommes. Si la répartition effective des hommes et des femmes dans une population est de 51 % de femmes et de 49 % d'hommes, cet échantillon sous-représente les femmes. Un bon nombre de caractéristiques de la population active sont liées au sexe. Par exemple, une plus grande proportion d'hommes ont des emplois. En rajustant les sous-poids pour tenir compte de la véritable proportion d'hommes et de femmes, on obtient une meilleure estimation. Le facteur de compensation déterminé pour tirer parti des données auxiliaires est appelé *facteur g*. Le produit du sous-poids et du facteur g est appelé *poids final*.

Pour obtenir le facteur g, l'EPA utilise une variante de l'estimateur de régression généralisé (ERG) fondée sur la méthode de pondération proposée par Lemaître et Dufour (1987). On utilise, comme données auxiliaires, des estimations postcensitaires de population projetées sur la période courante. Plus précisément, on utilise les chiffres de population de 30 groupes d'âge-sexe dans chacune des provinces, ainsi que les chiffres de population des régions économiques et des régions métropolitaines de recensement. Ces chiffres de population sont produits chaque mois par la Division de l'analyse des enquêtes sur le travail et les ménages de Statistique Canada.

désiré. Ce facteur de rajustement est appelé le *poids de stabilisation*.

Il faut d'abord définir les secteurs de stabilisation. Dans le présent plan de sondage, un *secteur de stabilisation* est défini comme étant tous les logements qui appartiennent à la même REAE et au même groupe de renouvellement. On détermine la taille d'échantillon de base de chaque secteur de stabilisation **a**, c'est-à-dire quel échantillon on veut obtenir en fonction de la répartition de l'échantillon. Soit $\mathbf{b}_a$ la taille d'échantillon de base du secteur **a**. Si on effectue un échantillonnage sans stabilisation, un nombre donné de logements est sélectionné. Appelons ce nombre $\mathbf{n}_a$. Si la valeur $\mathbf{n}_a$ dépasse la valeur $\mathbf{b}_a$, il faut retirer $\mathbf{n}_a$ - $\mathbf{b}_a$ logements. Cette opération est effectuée systématiquement au hasard. Ensuite, il faut rajuster le poids de base.

L'EPA prévoit que si une grappe a été sous-échantillonnée par l'application de la méthode III (voir la section précédente), celle-ci doit être exclue du processus de stabilisation. Aucun logement de cette grappe ne peut être retiré et le poids de stabilisation n'est pas appliqué. Soit $\mathbf{c}_a$ le nombre total de logements du secteur de stabilisation **a** exclus de cette façon.

Il existe deux autres cas où on n'attribue aucun poids de stabilisation à un logement d'un secteur de stabilisation. À l'occasion, on découvre un groupe de logements qui n'étaient censés être qu'un seul logement. Ces logements, appelés *multiples*, sont tous inclus dans l'échantillon. Comme il n'y avait aucune chance qu'ils puissent être retirés par stabilisation, aucun poids de stabilisation ne leur est attribué. Aussi, durant l'existence d'une grappe, de nouveaux logements sont construits et s'ajoutent à la liste de ménages de la grappe. À nouveau, comme ils n'ont pas eu la chance d'être retirés, aucun poids de stabilisation n'est appliqué.

Une fois que les logements ont été retirés, le secteur de stabilisation est subdivisé en sous-secteurs. Un sous-secteur de stabilisation est un groupe de strates du secteur de stabilisation qui possèdent la même fraction de sondage inverse $\mathbf{R}_h$. Les poids de stabilisation de chaque sous-secteur sont calculés séparément. Ce point subtil n'est pas indiqué dans la notation.

Le poids de stabilisation applicable aux ménages du secteur **a** est le suivant :

$$s_a = \frac{n_a - c_a}{b_a - c_a} .$$

Pour conclure cette section, nous rappelons que le poids de sondage, ou probabilité d'inclusion inverse, **w**, est défini comme étant le produit du poids de base $\mathbf{R}_h$, du sous-poids de grappe **K** et du poids de stabilisation $\mathbf{s}_a$.

Traitement de la non-réponse et calcul du sous-poids

Comme dans toutes les enquêtes, l'EPA enregistre de la non-réponse. La non-réponse est classée en deux catégories :

1. Il y a *non-réponse à une question* lorsqu'il manque un renseignement au sujet d'un ménage. Cela signifie qu'il manque certains détails mais non l'ensemble des détails relatifs à un ou plusieurs membres du ménage ou qu'il manque tous les détails relatifs à certains membres du ménage mais non à l'ensemble d'entre eux.
2. Il y a *non-réponse de l'unité* lorsqu'il n'existe aucune information sur aucun des membres du ménage.

La non-réponse à une question est traitée entièrement par imputation. En cas de non-réponse à une question en particulier, on trouve un enregistrement donneur parmi ceux fournis par les répondants. On prend la réponse du donneur qui correspond à l'information manquante. En général, un donneur acceptable est une personne qui possède des caractéristiques démographiques semblables dans un même secteur géographique et qui, aux questions pour lesquelles il existe une réponse, offre un schéma de réponse comparable. La méthode d'imputation est expliquée en détail dans Lorenz (1995).

Dans le cas de non-réponse du ménage (ou unité), si un ménage non-répondant a répondu le mois précédent, les réponses du mois précédent sont transférées. Cette méthode ne s'applique que s'il y a eu une réponse le mois précédent (c.-à-d. que les réponses ne sont pas transférées une deuxième fois).

Enfin, toutes les non-réponses d'une unité complète sont traitées par la méthode de rajustement des poids. Le rajustement des poids est fondé sur le principe que les ménages répondants peuvent être utilisés pour représenter tous les ménages répondants ou non répondants. Le poids de sondage est multiplié par ce facteur de compensation de la non-réponse (défini ci-après) et le résultat est appelé *sous-poids*.

Afin de procéder à ce rajustement des poids, l'échantillon est d'abord subdivisé en classes de rajustement de poids ou *secteurs de non-réponse*. Les secteurs de non-réponse sont définis de manière à augmenter les chances que les

$$R^{N}_{(hj)} = \frac{N^{N}_{(hj)}}{n^{N}_{(hj)}}$$

La valeur $\mathbf{R}^{N}_{(hj)}$ est le poids de base à attribuer aux ménages à sélectionner dans cette nouvelle strate. Comme le poids $\mathbf{R}_h$ de l'ancienne strate est attribué aux ménages sélectionnés dans cette nouvelle strate, le facteur approprié est le suivant :

$$K = \frac{R^{N}_{(hj)}}{R_h}$$

Il est également nécessaire d'appliquer un facteur de rajustement à tous les autres ménages échantillonnés du reste de la strate originale. Prenons une strate dans laquelle six grappes ont été sélectionnées. Les six grappes représentent l'ensemble de la population de la strate. Une fois que la grappe en croissance a été retirée de la strate, il faut rajuster le poids des ménages des cinq grappes restantes afin qu'elles soient représentatives du reste de la strate.

Soit $\mathbf{N}^{R}_{h} = \mathbf{N}_h - \mathbf{N}^{N}_{(hj)}$ le compte d'origine du reste de la strate et, si $\mathbf{n}_{hj}$ est le rendement prévu initial de la grappe qui a été retirée de la strate, soit $\mathbf{n}^{R}_{h} = \mathbf{n}_h - \mathbf{n}_{hj}$ le rendement prévu de l'échantillon du reste de la strate. La nouvelle fraction de sondage inverse de la strate est la suivante :

$$R^{R}_{h} = \frac{N^{R}_{h}}{n^{R}_{h}}$$

On obtient alors le sous-poids de grappe suivant :

$$K = \frac{R^{R}_{h}}{R_h}$$

Méthode III : Sous-échantillonnage de grappe

Lorsqu'une grappe doit être sous-échantillonnée et que les méthodes I ou II ne conviennent pas, on a recours à cette méthode qui est la façon la plus simple et la plus courante pour obtenir un sous-échantillon. La grappe est d'abord échantillonnée en appliquant les anciennes fractions de sondage. On obtient un premier échantillon de ménages. Un deuxième choix au hasard est effectué parmi les ménages sélectionnés. Les ménages retenus après la deuxième sélection sont interviewés, tandis que les autres sont retirés de l'échantillon. Si la grappe a initialement été échantillonnée en appliquant l'ancienne fraction de sondage $\mathbf{R}_{hj}$ et que le sous-échantillonnage donne une fraction de sondage $\mathbf{R}^{*}_{hj}$, alors le sous-poids de la grappe est le suivant :

$$K = \frac{R^{*}_{hj}}{R_{hj}}$$

Par exemple, si on retient un ménage sur deux dans l'échantillon, la nouvelle fraction de sondage est égale au double de l'ancienne fraction de sondage de la grappe. La fraction ci-dessus serait égale à deux. Le poids de base des ménages de cette grappe sera multiplié par deux pour compenser les ménages abandonnés. À cause des résultats aberrants que peuvent donner les enquêtes spéciales effectuées au moyen du plan de sondage de l'EPA, le sous-poids de grappe ne peut pas avoir une valeur supérieure à 3.

Poids de stabilisation

Au dernier degré d'échantillonnage, on effectue un échantillonnage systématique à un taux fixe. Comme on emploie toujours la même fraction de sondage, la croissance de la population et, par voie de conséquence, du nombre de ménages entraîne un élargissement progressif de l'échantillon et une augmentation des coûts de sondage. Afin de limiter les coûts, on procède à une stabilisation de l'échantillon. La stabilisation d'un échantillon est le retrait au hasard de logements de l'échantillon dans le but de maintenir celui-ci au niveau désiré. En retirant des logements au hasard, on se trouve à modifier la probabilité d'inclusion des ménages. Par exemple, supposons la définition d'un secteur de stabilisation **a** comprenant des ménages dont la probabilité d'inclusion était de 1 sur 200 au moment de l'établissement du plan de sondage. Si le secteur de stabilisation a un rendement prévu de 300 logements et que l'échantillonnage selon la probabilité attribuée donne 350 logements, il faut retirer 50 logements. Après le retrait des 50 logements, la probabilité d'inclusion n'est plus de 1 sur 200, mais de 3 sur 700 (c.-à-d. 1/200 multiplié par 300/350). Comme dans le sous-échantillonnage de grappe, il est plus facile de rajuster les poids de base au besoin que de continuellement les recalculer. Le poids de base reste à 200, mais il est multiplié par le facteur 350/300 pour obtenir le poids

La probabilité conditionnelle de sélection du ménage **k** étant donné que l'UPD **j** a été sélectionnée est, par définition, la suivante :

$$\pi_{k/j} = \frac{n^*_{hj}}{N_{hj}} = \frac{1}{R_{hj}}$$

La probabilité d'inclusion du ménage **k** dans la strate **h** est alors la suivante :

$$\pi_{hk} = \pi_{1hj}\,\pi_{k/j} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}}\,R_{hj}\,\frac{1}{R_{hj}} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}}$$

Il est à noter que :

$$\sum_{j\in h} R_{hj} = \sum_{j\in h} \frac{N_{hj}}{n^*_{hj}} = \frac{n_{1h}}{n_h} \sum_{j\in h} N_{hj} = n_{1h}R_h$$

La probabilité d'inclusion est égale à la FSI initiale de la strate, à savoir $\mathbf{1/R_h}$. En général, un plan de sondage dont les poids de base sont égaux dans chaque strate est appelé un *plan de sondage autopondéré*. L'EPA est autopondérée dans chaque strate (en ce qui concerne le poids de base), et le poids de base correspond à $\mathbf{R_h}$.

Sous-poids de grappe

Comme nous l'avons déjà mentionné, l'EPA comporte un plan de sondage à plusieurs degrés. Les avant-dernières unités, à savoir les grappes, sont échantillonnées à un taux fixe déterminé en fonction des comptes du recensement de 1991, de manière à obtenir un rendement de 6 à 10 logements par grappe. Dans les secteurs urbains, les nouvelles expansions sont fréquentes et le nombre de logements peut s'accroître considérablement. À ce moment, comme la fraction de sondage est fixe, la tâche de l'intervieweur peut augmenter considérablement, ce qui peut avoir des répercussions sur la qualité de son travail et sur sa capacité de mener à bien sa tâche. Lorsque la croissance d'une grappe dépasse 200 %, celle-ci peut être sous-échantillonnée selon l'une des trois méthodes décrites ci-après. Dans chacune d'entre elles, un certain nombre de ménages est retiré au hasard de la grappe où la population s'est accrue, ce qui a pour effet de modifier la probabilité d'inclusion du ménage. Au lieu de constamment recalculer le poids de base, il est plus simple de calculer un facteur de rajustement du poids et de l'appliquer au poids de base. Ce facteur de rajustement est appelé *sous-poids de grappe*. Les trois méthodes utilisées pour retirer des ménages et déterminer le sous-poids de grappe sont les suivantes :

Méthode I: Formation de sous-grappes

Lorsque la croissance dépasse 300 % et que les quadrilatères sont assez bien définis pour délimiter des grappes, la grappe en croissance est subdivisée en sous-grappes. Un échantillon des plus petites grappes est pris, soit $\mathbf{n_{2hj}}$ d'entre elles. Les plus petites grappes sont échantillonnées de manière à réduire le rendement global. Soit $\mathbf{R_{hj}}$ une valeur égale à la fraction de sondage de la grappe originale. La taille $\mathbf{N_{hji}}$ des nouvelles grappes et le rendement prévu $\mathbf{n_{hji}}$ de l'échantillon donnent les fractions de sondage des nouvelles grappes, à savoir $\mathbf{R_{hji}}$. On détermine ensuite quelle doit être la fraction de sondage de l'ancienne grappe pour obtenir la taille de l'échantillon global prélevé des nouvelle sous-grappes. Cette valeur est fournie par la formule :

$$R^*_{hj} = \sum_{i\in j} \frac{R_{hji}}{n_{2hj}}$$

Le sous-poids de la grappe est fourni par la formule :

$$K = \frac{R^*_{hj}}{R_{hj}}$$

On multiplie par ce facteur le poids de base attribué initialement aux ménages sélectionnés pour indiquer leur nouvelle probabilité de sélection.

Méthode II: Grappe autoreprésentative

Lorsque les caractéristiques des secteurs en croissance sont différentes de celles du reste de la strate ou que la taille de la grappe équivaut à au moins 20 % de celle d'une strate, la grappe est d'abord redéfinie en tant que strate. Soit la nouvelle strate **(hj)** . On forme de nouvelles grappes à l'intérieur de cette nouvelle strate et on prélève un échantillon. L'échantillon prélevé de la grappe en croissance est alors représentatif de la grappe elle-même et non de la strate originale de plus grande taille. Si le compte d'origine de la nouvelle strate est de $\mathbf{N^N_{(hj)}}$ et que le rendement prévu de l'échantillon est de $\mathbf{n^N_{(hj)}}$, la fraction de sondage de la strate est fournie par la formule :

de sondage et qu'il sera nécessaire de modifier les procédures d'échantillonnage et d'estimation.

4) Au moment de la conception de l'enquête, on se sert des données du dernier recensement. Dans ce cas-ci, les comptes d'origine (par exemple, le nombre de ménages dans un îlot) proviennent du recensement de 1991.

En fonction de la répartition de l'échantillon, le plan de sondage détermine un premier ensemble de poids de sondage. Ces poids pourraient être utilisés tant que le plan de sondage et que la répartition ne seraient pas modifiés. Cependant, étant donné que les unités de l'avant-dernier échelon sont appelées à croître et que le taux de l'échantillonnage systématique est fixe, la taille de l'échantillon est portée à augmenter continuellement (ainsi que le coût de la collecte des données). Ce phénomène entraîne également de grandes disparités dans la charge de travail des intervieweurs, dans une même charge de travail sur une période donnée et entre diverses charges de travail. Pour y parer, on a recours à deux méthodes de contrôle de la taille de l'échantillon. Ces deux méthodes : la stabilisation de l'échantillon et le sous-échantillonnage de grappe (décrites ci-après), modifient la probabilité d'inclusion d'un ménage dans l'échantillon. Il est donc nécessaire, pour compenser, de rajuster les poids de sondage attribués au départ. Les facteurs de compensation sont désignés sous le nom de *poids de stabilisation* et de *sous-poids de grappe*.

Dans les méthodes de stabilisation de l'échantillon et de sous-échantillonnage de grappe, le problème de la croissance de l'échantillon est résolu par le retrait d'un certain nombre de ménages. La stabilisation compense l'accroissement graduel résultant de la croissance de la population, qui, si rien n'est fait, entraîne l'élargissement de l'échantillon. Quant au sous-échantillonnage de grappe, il compense la croissance circonscrite à certains secteurs relativement restreints, qui peut affecter la charge de travail des intervieweurs.

Dans le présent document, les poids de sondage sont désignés sous le nom de *poids de base*. Le terme *poids de sondage* est réservé pour désigner la probabilité inverse d'inclusion qui s'applique au moment de l'échantillonnage. Habituellement, les termes «poids de sondage» et «poids de base» sont utilisés de manière interchangeable. Dans le cas présent, cette distinction ne sert qu'à marquer cette différence.

En résumé, le poids de sondage d'un ménage donné est égal à la probabilité inverse de son inclusion dans l'échantillon. Dans l'Enquête sur la population active, il est le produit de trois facteurs, à savoir : le poids de base, le sous-poids de grappe et le poids de stabilisation.

Poids de base

Au moment de la conception de l'enquête, on a formé des strates en regroupant des unités géographiques. La stratification est décrite au chapitre 2. Le nombre de ménages à sélectionner dans chaque strate est déterminé et fixe. Soit $\mathbf{n}_h$, le nombre de ménages sélectionnés pour la strate **h**. Nous connaissons également le nombre de ménages, $\mathbf{N}_h$, compris dans la strate au moment de la conception de l'enquête. La fraction de sondage inverse (FSI) de la strate est fournie par la formule :

$$R_h = \frac{N_h}{n_h}$$

Comme l'EPA comporte un échantillonnage à plusieurs degrés, il est nécessaire de déterminer le nombre d'unités à sélectionner à chaque degré. Prenons le cas d'un échantillonnage à deux degrés. En fonction des comptes d'origine de chaque unité du premier degré (UPD), le rendement prévu de l'échantillon est fixe. Ce rapport est appelé *facteur de densité*, et pour l'UPD **j** de la strate **h**, il sera désigné par la valeur $\mathbf{n^*}_{hj}$. Le nombre $\mathbf{n}_{1h}$ d'UPD à sélectionner est fourni par le rapport $\mathbf{n}_h/\mathbf{n^*}_{hj}$. Si $\mathbf{N}_{hj}$ est le nombre de ménages dans l'UPD **j** de la strate **h**, la fraction de sondage de l'UPD est égale à $\mathbf{N}_{hj}/\mathbf{n^*}_{hj}$. Ce rapport est désigné par la valeur $\mathbf{R}_{hj}$. La valeur $\mathbf{R}_{hj}$ correspond au pas de sondage qui est utilisé pour la sélection systématique des logements au dernier degré d'échantillonnage. Il est parfois désigné sous le nom de FSI de grappe.

Dans certains cas, la valeur $\mathbf{n}_{1h}$ est fixe et la valeur $\mathbf{n^*}_{hj}$ est déterminée par le rapport $\mathbf{n}_h/\mathbf{n^*}_{1h}$. Dans chaque cas, la taille de la strate est fixée de façon à obtenir des échantillons de la grandeur voulue.

Nous pouvons maintenant déterminer que la probabilité d'inclusion d'un ménage est le produit des probabilités de sélection à chaque degré. La valeur $\mathbf{R}_{hj}$ est utilisée comme mesure de la taille dans l'échantillonnage PPT de la $\mathbf{j}^{\text{ième}}$ UPD de la strate **h**. La probabilité d'inclusion au premier degré de l'UPD **j** est la suivante :

$$\pi_{1hj} = \frac{n_{1h}}{\sum_{j \in h} R_{hj}} R_{hj}$$

CHAPITRE 5 : Pondération et estimation

Introduction

Les estimations sont produites d'après les données d'échantillon en tenant compte des caractéristiques connues du plan de sondage et en employant des techniques d'estimation fondées sur la théorie de l'échantillonnage d'enquête. Un poids d'enquête (ou poids final) est attribué à chaque personne comprise dans l'échantillon. Ce poids correspond au nombre de personnes représentées par le répondant dans l'ensemble de la population et sert à obtenir des estimations pour toutes les caractéristiques à l'étude. Il est le produit de trois facteurs: un poids de sondage, qui incorpore les données relatives au plan de sondage; une compensation de la non-réponse, qui tient compte des ménages non répondants et un facteur (le facteur g), qui rajuste l'échantillon en fonction des chiffres de population connus.

Après avoir obtenu les estimations, il faut en déterminer la fiabilité. Étant donné le caractère probabiliste de l'échantillon de l'EPA, il est possible d'évaluer l'erreur d'échantillonnage correspondant à chaque estimation. L'erreur d'échantillonnage permet d'énoncer certains faits liés à la probabilité à propos des estimations de l'enquête.

Occasionnellement, il est nécessaire d'obtenir des estimations pour des régions non prévues au moment de la conception de l'enquête ou dont les limites ont été modifiées par la suite. Tel est le cas des REAE. Dans le cadre de son mandat d'administration du programme d'assurance-emploi canadien, DRHC doit obtenir des estimations pour ces régions. En utilisant des techniques d'estimation adaptées aux petites régions, il est possible d'améliorer la qualité des données relatives à ces régions.

Le présent chapitre a pour but de fournir une description des méthodes d'estimation de l'EPA et d'indiquer les raisons pour lesquelles ces méthodes ont été utilisées. Suit une description de la méthode d'estimation de l'erreur d'échantillonnage. Une section est consacrée à la méthode de dérivation des estimations pour les REAE. Une dernière section traite des changements méthodologiques entre l'ancien et le nouveau plan de sondage et des données auxiliaires qui entrent dans la pondération.

Poids de sondage

Dans toute enquête par sondage, une population-cible est définie. Celle-ci constitue un sous-ensemble de la population possédant les caractéristiques à étudier. Dans tout échantillon, certains membres de la population sont sélectionnés et d'autres ne le sont pas. Les membres sélectionnés peuvent être considérés comme étant représentatifs de ceux qui n'ont pas été sélectionnés. Dans un échantillon probabiliste, chaque membre a une probabilité connue d'être sélectionné. Si cette probabilité de sélection est, disons, de un sur cinquante, alors le membre représente en tout 50 personnes. On pourrait faire cinquante copies des réponses au sondage et, en répétant cette opération pour chaque membre de l'échantillon, créer une «pseudo-population». On pourrait obtenir les estimations voulues à partir de cette pseudo-population, car, si l'échantillon est représentatif de la population, les calculs portant sur celle-ci constitueront une bonne approximation des résultats qui auraient pu être obtenus en prenant l'ensemble de la population. Dans les faits, au lieu de copier les enregistrements, on leur attribue un poids. Comme ce poids est déterminé par le plan de sondage, il est appelé «poids de sondage». Le poids de sondage peut être considéré comme le nombre de fois que l'enregistrement aurait été copié.

En ce qui concerne l'EPA, la méthode d'estimation est conditionnée par les faits suivants :

1) L'enquête est constituée suivant un plan de sondage stratifié à plusieurs degrés comportant un échantillonnage avec PPT à toutes les étapes, sauf à la dernière, où l'échantillonnage est systématique.
2) Comme les unités d'échantillonnage sont des ménages, les poids de sondage, dans l'EPA, se rapportent à des ménages. Comme nous l'avons déjà mentionné, les données recueillies se rapportent à tous les membres du ménage. Afin de pouvoir obtenir des estimations relatives aux personnes, le poids de sondage relatif au ménage est attribué à chacun des membres du ménage.
3) L'EPA est une enquête périodique. Une fois qu'il a été établi, le même plan de sondage est repris d'un mois à l'autre jusqu'à ce qu'un nouveau plan de sondage soit adopté. Depuis qu'elle existe, l'enquête a été remaniée tous les dix ans. Il est prévu que la population s'accroîtra pendant la durée de vie du plan

On choisit un membre du ménage initialement sélectionné pour participer à une interview détaillée et on le suit pendant une période prévue de 20 ans, à raison d'une interview tous les deux ans. À chaque cycle, on recueille, à des fins d'estimations transversales, des renseignements de base sur la santé de tous les membres du ménage qui résident alors avec le répondant longitudinal.

On trouvera une description plus détaillée de la méthodologie de l'ENSP dans Tambay et Catlin (1995).

L'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), qui a été lancée en 1994, porte sur un échantillon d'enfants dont elle suit le développement pendant un certain nombre d'années, de la petite enfance à l'âge adulte. Il s'agit d'une enquête complexe, qui, à l'origine, s'est servie de l'échantillon des ménages de l'EPA pour former un échantillon d'enfants et pour obtenir ensuite des renseignements de la part des enseignants et des directeurs des établissements que fréquentent ces enfants. Comme seulement environ 30 % des ménages de l'EPA ont des enfants ayant l'âge requis pour l'ELNEJ, on a dû recourir à plus de six groupes de renouvellement de l'EPA pour former l'échantillon nécessaire. Dans la plupart des cas, on a eu recours à huit ou neuf groupes de renouvellement. Outre l'échantillon provenant directement de l'EPA, l'ELNEJ comprend également les enfants provenant de 2 500 ménages sélectionnés lors du premier cycle de l'ENSP. L'échantillon direct de l'EPA et celui de l'ENSP représentent respectivement 21 000 et 4 000 enfants de moins de 12 ans, pour un échantillon total initial de 25 000 individus. Tous les deux ans, les personnes échantillonnées sont interviewées, et à cette occasion on augmente l'échantillon d'enfants dans les échantillonnées groupes qui ne sont plus représentés dans l'échantillon initial.

Dans le cadre de l'ELNEJ, on utilise une variété de questionnaires et de méthodes de collecte. La première interview a lieu en personne au moyen du mode d'interview assistée par ordinateur. Les enfants de 10 et 11 ans répondent à un questionnaire à remplir soi-même. L'enseignant et le directeur d'école de chaque enfant sont identifiés et on leur demande de remplir un questionnaire. Dans ces derniers cas, on procède par la technique d'envoi et de retour par la poste. Outre les types habituels de questionnaire, on administre également deux tests aux enfants, une épreuve de mathématiques et un test de compréhension de vocabulaire. On trouvera de plus amples renseignements sur l'ELNEJ dans le document de Brodeur et coll. (1995).

Ces résultats ont été utilisés pour dériver des mesures reliées au revenu comme par exemple les seuils de faible revenu. On songe présentement à intégrer l'EFC à l'Enquête sur la dynamique du travail et du revenu (EDTR). Une discussion suit ci-bas.

L'Enquête sur les dépenses des ménages (EDM) est une nouvelle enquête annuelle auprès des ménages qui remplace l'Enquête sur les dépenses des familles (EDF) et l'Enquête sur les dépenses alimentaires. Cette nouvelle enquête est introduite dans le cadre du Projet d'amélioration des statistiques économiques provinciales (PASEP). L'objectif général du PASEP est d'obtenir des données provinciales plus fiables pour les besoins de la formule de répartition fiscale pour la taxe de vente harmonisée. L'Enquête sur les dépenses des ménages continuera de remplir sa fonction traditionnelle de source d'information pour le calcul de l'indice des prix à la consommation. L'EDM est une enquête spéciale, ce qui implique que les ménages de cette enquête sont choisis dans les grappes échantillonnées par l'EPA, mais les ménages faisant partie de l'EDM ne sont pas interrogés dans le cadre de l'EPA.

La nouvelle enquête sera très différente de l'EDF, qui était menée tous les quatre ans. L'EDM aura lieu tous les ans et son échantillon sera deux fois plus important que celui de l'EDF. Par conséquent, l'EDM épuisera les grappes de l'EPA plus rapidement. La plus grande différence entre les deux enquêtes se rapportera à la méthode de collecte des données. Le long questionnaire de l'EDF sera remplacé par une approche simplifiée impliquant des contacts de nature variée avec les ménages répondants.

Enquêtes longitudinales. Dans les années 1990, Statistique Canada a mis au point différentes enquêtes longitudinales pour obtenir des données visant à combler certaines lacunes en matière d'information sur les Canadiens. Les principales enquêtes entrant dans cette catégorie sont l'Enquête sur la dynamique du travail et du revenu, l'Enquête longitudinale nationale sur les enfants et les jeunes et l'Enquête nationale sur la santé de la population. En voici un bref aperçu.

L'Enquête sur la dynamique du travail et du revenu (EDTR) a été créée pour étudier les processus qui influencent la vie économique des Canadiens. On utilise cette enquête pour observer les mouvements d'entrée et de sortie relativement à la situation de faible revenu, les transitions du marché du travail et la relation entre la dynamique familiale et le bien-être économique. Le premier panel de l'Enquête sur la dynamique du travail et du revenu a été constitué en 1993 et le deuxième en 1996. Les deux panels se chevauchent, et chaque panel demeure dans l'enquête pendant six ans. Le premier panel sera donc remplacé en 1999. À l'origine, chaque panel est formé de ménages ayant été récemment interviewés dans le cadre de l'EPA. Au cours de la durée de vie d'un panel, les individus qui en font partie pourront être interviewés jusqu'à 12 fois, passant d'une interview sur leurs activités professionnelles en janvier à des questions sur leur revenu en mai (les personnes peuvent éviter l'interview de mai en permettant à Statistique Canada d'utiliser leurs données fiscales de sources administratives). À l'instar d'autres enquêtes longitudinales, l'EDTR suit la situation des individus échantillonnés dans le temps, même s'ils déménagent dans une autre province ou à l'extérieur du pays.

Comme l'EDTR servira à produire des estimations tant transversales que longitudinales, l'échantillon des deux panels de longue durée sera complété par un échantillon de rafraîchissement pour garantir que l'échantillon total est représentatif de la population à un point donné dans le temps. Chaque échantillon de rafraîshissement est constitué de 10 000 ménages qui seront combinés à l'échantillon longitudinal (30 000 ménages) pour produire des estimations transversales.

On a lancé l'*Enquête nationale sur la santé de la population (ENSP)* pour évaluer dans le temps la santé des Canadiens. L'ENSP était la première enquête a utilisé le nouveau plan de sondage de l'EPA, à compter de juin 1994, par la sélection de 25 000 ménages. Ces ménages n'ont pas été interviewés dans le cadre de l'EPA. Pour répondre à des demandes spéciales, y compris l'achat d'échantillons par certains gouvernements provinciaux, l'ENSP complète l'échantillon provenant de la base de l'EPA par un système d'appels aléatoires. On rejoint ainsi 800 ménages supplémentaires au cours du premier cycle de l'enquête. Au cours du deuxième cycle, environ 60 000 ménages supplémentaires sont échantillonnés au moyen du système d'appels aléatoires afin de produire des données infra-provinciales à l'intention de trois provinces.

Pour tenir compte des facteurs saisonniers, on répartit l'échantillon de l'ENSP sur quatre trimestres, les interviews étant réalisées en février, en juin, en août et en novembre.

CHAPITRE 4 - Enquêtes spéciales et enquêtes supplémentaires

On se sert de la base et de l'échantillon de l'Enquête sur la population active pour recueillir de l'information dans le cadre de nombreuses enquêtes auprès des ménages. On appelle *enquêtes supplémentaires* les enquêtes pour lesquelles on interroge les ménages qui ont été sélectionnés dans le cadre de l'EPA. Les enquêtes qui utilisent la base de l'EPA pour choisir un échantillon différent de ménages sont appelées *enquêtes spéciales*. Dans ces derniers cas, les ménages sont habituellement sélectionnés des grappes qui servent également pour les interviews de l'EPA. Ces pratiques permettent de réaliser des économies de coûts considérables. Les enquêtes spéciales et supplémentaires sont souvent parrainées par d'autres ministères ou organismes publics. Soulignons que les enquêtes supplémentaires peuvent être réparties en deux groupes : celles qui interrogent les ménages de l'EPA alors qu'ils sont toujours sondés dans le cadre de l'EPA et celles qui interrogent des ménages qui ne font plus partie de l'EPA.

On peut utiliser chacun des six groupes de renouvellement de l'EPA pour produire des estimations. Normalement, les enquêtes spéciales et les enquêtes supplémentaires utilisent de un à cinq groupes de renouvellement pour établir leur échantillon, selon le niveau de fiabilité nécessaire. Habituellement, on évite le groupe de départ de l'EPA, c'est-à-dire celui formé de ménages sondés pour la première fois dans le cadre de l'EPA, parce que la première interview de l'EPA prend plus de temps à réaliser que les interviews suivantes.

Dans certains cas, on n'a besoin que d'une portion d'un groupe de renouvellement. Pour ce faire, on retire des ménages au hasard comme dans le programme de stabilisation de l'EPA. On peut également procéder à une sélection parmi les ménages en prélevant un échantillon aléatoire ou en éliminant les individus selon certaines caractéristiques.

Le tableau suivant dresse la liste de certaines des enquêtes ayant utilisé les groupes de renouvellement de l'EPA ou la base de sondage de l'EPA en 1998.

Exemples des principales enquêtes spéciales et enquêtes supplémentaires

Enquêtes	Période de collecte de données
Enquête sur les voyages des Canadiens	Janvier-décembre (mensuelle)
Enquête sur la couverture de la population pour le Programme d'assurance-emploi	Janvier
Enquête sur les dépenses des ménages	Janvier - mars
Enquête sur la dynamique du travail et du revenu	Janvier et mai
Enquête sur l'éducation et sur la formation des adultes	Janvier
Enquête sur le service téléphonique résidentiel	Février, mai, août, novembre
Enquête nationale sur la consommation d'énergie des ménages	Février
Enquête sur les réparations et les rénovations effectuées par les propriétaires-occupants au Canada	Mars
Enquête sur les finances des consommateurs	Avril
Enquête sur le patrimoine culturel	Avril
Enquête nationale sur la santé de la population	Juin, août, novembre (février 1999)
Enquête longitudinale nationale sur les enfants et les jeunes	Novembre
Enquête sur les horaires et les conditions de travail	Novembre

L'Enquête sur les finances des consommateurs (EFC) est une enquête auprès des ménages réalisée en avril. Cette enquête est un supplément de l'EPA qui utilise les ménages de quatre groupes de renouvellement. Un questionnaire est posté à chacun de ces ménages avant l'interview de l'EPA d'avril. L'information est alors recueillie au moment de l'interview de l'EPA au moyen de l'interview assistée par ordinateur. Les principaux extrants de l'EFC comprennent les distributions de revenu avant et après impôt et les revenus moyens et médians.

Développement des ressources humaines. En 1996, les 61 RAC ont été remplacées par 53 régions économiques d'assurance-emploi. Une fois arrêtées les limites de ces nouvelles régions, on a étudié la taille de l'échantillon par région et la qualité des estimations obtenues. Dans cinq cas, l'échantillon affecté à une région était trop faible pour répondre aux exigences du Programme d'assurance-emploi. On a donc dû augmenter l'échantillon dans ces régions en 1997, et procéder à une diminution correspondante de la taille de l'échantillon dans les régions qui pouvaient supporter une telle réduction. À l'échelle nationale, la taille de l'échantillon national est demeurée identique, mais il y a eu des échanges entre les provinces. Le tableau A0 de l'annexe A donne un aperçu de l'EPA après la mise en place de ces changements.

lorsque l'on veut obtenir des sous-échantillons de l'EPA.

Pour atteindre cet objectif, on a attribué des numéros de renouvellement aux grappes échantillonnées de manière à ce que le rendement total prévu soit le même dans chaque strate et région économique d'assurance-emploi. L'attribution des numéros de renouvellement s'est faite de façon indépendante dans chaque REAE, mais on a introduit un facteur de l'origine aléatoire permettant de répartir le rendement prévu aussi également que possible aux niveaux plus globaux.

Dans la plupart des secteurs, chaque strate comporte six sélections ou un multiple de six sélections, de sorte que chaque groupe de renouvellement peut avoir le même nombre de sélections. Dans les secteurs ruraux où un plan à trois degrés est utilisé, on forme de six à neuf grappes, de sorte que l'on doit regrouper certaines strates pour former six groupes de renouvellement. En général, on combine les unités plus petites pour égaliser plus ou mois la répartition. La base d'appartements comporte un nombre variable de sélections qui reçoivent au hasard un numéro de renouvellement. Normalement, la base des secteurs éloignés comporte moins de six sélections et fait l'objet d'un traitement à part - leur rendement réel est, en tout cas, très incertain.

Comme la taille et le nombre d'unités échantillonnées par strate ne sont pas les mêmes dans les bases d'appartements, les secteurs urbains et les secteurs ruraux, la première chose à faire pour uniformiser le rendement de l'échantillon à chaque renouvellement a été d'essayer de faire en sorte que celui-ci soit aussi uniforme que possible dans les grandes unités rurales et dans les REAE. Des numéros de renouvellement ont ensuite été attribués aux unités d'appartements, qui sont très variables, et ensuite aux unités urbaines pour uniformiser le rendement au niveau régional.

Essentiellement, l'attribution est effectuée au moyen d'un tableau hiérarchisé des rendements prévus par rotation. Pour chaque strate figurant dans la liste d'une REAE, les rendements de renouvellement sont classés du plus faible au plus élevé. Le total cumulatif du rendement obtenu jusqu'à maintenant est trié également mais en ordre inverse. On procède au renouvellement en appariant le renouvellement comptant le plus bas rendement dans la strate avec celui ayant le plus haut total cumulatif. Au début, on donne aux renouvellements un faible rendement aléatoire. À la fin de la liste, les rendements ne devraient pas varier plus que la variation à l'intérieur de n'importe quelle strate. Notons qu'il s'agit là de rendements prévus, et que, dans certains cas, l'échantillon réel s'écartera considérablement de cette valeur prévue.

Comme la base d'appartements est ouverte, il faut continuellement lui attribuer des numéros de renouvellement. On attribue à chaque ensemble de six sélections de l'échantillon systématique les six numéros de renouvellement selon un ordre aléatoire. Il n'est pas possible d'uniformiser le rendement de l'échantillon dans cette base ouverte. Pour les nouvelles unités qui y sont ajoutées, on produit au besoin une série de six numéros de renouvellement aléatoires. Pour attribuer les numéros de renouvellement, il suffit de passer au numéro suivant de la série.

De même, pour les strates à trois degrés d'échantillonnage, l'étape de sélection au niveau de la grappe n'est pas terminée tant que l'UPE fait la rotation dans l'échantillon. La nouvelle UPE doit recevoir un numéro de renouvellement au moment de l'introduction. Le reste du rendement est calculé sans tenir compte de l'échantillon en question. On utilise ensuite la même technique pour attribuer les numéros à la nouvelle UPE.

Il arrive à l'occasion qu'une grappe reçoive un numéro de renouvellement qui ne suit pas le tableau de dates d'introduction établies selon les règles. On parle alors de sélections hors renouvellement. C'est dans la base d'appartements que cela arrive le plus souvent. Comme elle est ouverte, une nouvelle grappe peut être sélectionnée chaque mois. Simultanément, le numéro de renouvellement attribué est aléatoire. Plutôt que d'attendre les mois 2 à 5 pour introduire l'échantillon dans le cycle de renouvellement, il est préférable que la grappe soit traitée pour interview le plus tôt possible, d'où la nécessité d'une sélection hors renouvellement.

Changements apportés à l'Enquête sur la population active après le remaniement

Après la mise en place du nouvel échantillon de l'EPA en mars 1995, deux modifications d'importance ont été apportées à l'enquête. La première est une réduction de 11 % de la taille de l'échantillon à compter de juillet 1995. Comme le nouveau plan de l'EPA est plus efficient que le précédent, les coefficients de variation des estimations nationales et provinciales se comparent à ceux qui existaient avant le remaniement. L'échantillon a été réduit dans les mêmes proportions dans chaque province.

Le deuxième changement d'importance est la redéfinition des régions d'assurance-chômage par le ministère du

Processus de renouvellement de l'échantillon. Le renouvellement de l'échantillon se fait automatiquement au moyen du Système de conception de l'échantillon, qui détermine quelles unités doivent être introduites dans l'échantillon de chaque enquête et lesquelles doivent en être supprimées. Ces unités sont traitées manuellement par le personnel de la Section du contrôle de l'échantillon, qui doit déterminer quelles régions géographiques sont représentées et former, s'il y a lieu, les unités des degrés suivants d'échantillonnage.

Chaque type différent de base dans le plan d'échantillon utilise des programmes de sélection de grappes qui créent des enregistrements de renouvellement. Ces enregistrements indiquent l'ordre dans lequel on ajoute 1 aux origines aléatoires à l'intérieur des grappes et l'ordre dans lequel les grappes prélevées dans un groupe se succèdent. Chaque enregistrement est valide tant que le groupe est soumis à l'échantillonnage, jusqu'à concurrence de 40 origines aléatoires. L'ensemble de ces enregistrements constituent le fichier principal de renouvellement, qui sert au renouvellement automatisé de l'échantillon. Le fichier est modifié chaque fois que les plans de renouvellement sont mis à jour (comme pour la base ouverte d'appartements) et au fur et à mesure que les enregistrements deviennent périmés et doivent être remplacés.

Le renouvellement des ménages se fait de la manière suivante. Sept mois avant la date de l'enquête, les renseignements relatifs au plan de sondage de toutes les grappes devant subir un renouvellement pour cette date sont établis. Cela comprend les grappes nouvelles et les grappes existantes.

Pour les nouvelles grappes, la Section du contrôle élabore les diagrammes de grappes (F01) nécessaires. Ces diagrammes sont ensuite transmis aux bureaux régionaux (BR) de 20 à 23 semaines avant la date d'introduction afin d'établir la liste. On introduit également dans les ordinateurs portatifs les nouveaux fichiers de contrôle de liste qui coïncideront avec ces F01. L'intervieweur transmet à son tour à la base de données centrale les renseignements obtenus sur les ménages.

On a déjà introduit dans la base de données centrale l'origine aléatoire et la fraction de sondage inverse devant être appliquées à chaque grappe, nouvelle ou existante, de l'échantillon pour une enquête et une date particulière. On procède à une sélection des ménages en prélevant un échantillon systématique des listes environ six semaines avant la semaine d'interview. Les enregistrements sur les ménages sont retransmis aux ordinateurs portatifs des intervieweurs pour l'étape de l'interview. Ces ménages choisis demeurent dans l'échantillon pendant six mois.

Tous les six mois, une autre série d'origines est transmise à la base de données et de nouveaux ménages sont échantillonnés. Pendant ce temps, les origines des autres renouvellements sont également sélectionnés.

Dans le plan de sondage à trois degrés, la durée d'un enregistrement de renouvellement équivaut seulement à la période pendant laquelle l'UPE demeure dans l'échantillon. Ainsi, à mesure que les enregistrements s'épuisent, on peut déterminer les UPE à remplacer. En se reportant à la base du plan, on identifie les UPE de remplacement. La Section du contrôle de l'échantillon détermine alors l'emplacement géographique des UPE entrant dans le cycle et prépare les cartes du secteur.

En général, le processus de renouvellement des UPE débute au moins trente semaines avant la date d'introduction de l'échantillon. Il faut s'y prendre à l'avance puisque, dans ces nouveaux secteurs, certaines grappes n'ont pas encore été formées. Après le renouvellement des UPE, il faut procéder à l'échantillonnage des unités suivantes, soit les grappes, ce qui créera de nouveaux enregistrements de renouvellement.

Attribution des numéros de renouvellement. Le but poursuivi lorsque l'on attribue les numéros de renouvellement, c'est que le rendement prévu soit le même partout. Le rendement prévu est obtenu par l'échantillonnage de toutes les grappes d'après les comptes sur les ménages utilisés pour créer la base. On procède à cette distribution tant pour l'échantillon dans son ensemble que pour les plus petites unités géographiques de celui-ci. Si l'on respecte cet objectif, il en découle trois conséquences :

- la charge de travail des intervieweurs est stable, étant donné qu'un nombre à peu près égal d'unités est remplacé chaque mois ;

- l'échantillon est composé d'un nombre égal de ménages qui en font partie depuis un, deux, trois, quatre, cinq ou six mois, ce qui annule l'effet que le nombre des mois passés dans l'échantillon pourrait avoir sur les estimations relatives aux différents secteurs et à différentes périodes ;

- l'échantillon est bien divisé en six parties représentatives de même taille, dont on peut se servir

la grappe. Le premier détermine l'origine aléatoire pour l'échantillonnage systématique des logements dans la grappe. Le second détermine le nombre d'échantillons systématiques de logements à prélever dans la grappe, c'est-à-dire le nombre de périodes de six mois pendant lesquelles la grappe fera partie de l'échantillon.

Avant chaque prélèvement d'un nouvel échantillon de logements, on ajoute 1 à l'origine aléatoire de la grappe, jusqu'à ce que la valeur obtenue dépasse la FSI de la grappe ; à ce moment-là, l'origine redevient 1. Une fois que l'on a atteint le nombre de prélèvements de l'échantillon qui a été choisi au hasard, on procède au remplacement de la grappe.

Il est nécessaire d'attribuer à chaque première grappe choisie un nombre aléatoire de périodes pendant lesquelles elle restera dans l'échantillon pour maintenir la probabilité initiale de sélection des unités. Si, par exemple, on gardait les premières unités choisies jusqu'à épuisement de l'échantillon, c'est-à-dire jusqu'à ce que tous les échantillons systématiques de logements aient été prélevés, l'échantillon finirait par devenir biaisé, car les grandes unités y seraient surreprésentées. On procède au renouvellement en passant à la grappe suivante dans la liste randomisée de grappes par groupe. Si on est arrivé au bas de la liste, on revient à la première grappe de la liste. Comme on le fait pour choisir la première grappe, le choix des logements est déterminé par une origine aléatoire entre 1 et la FSI, à laquelle on ajoute 1 à chaque prélèvement de l'échantillon.

Secteurs où l'on applique la méthode de l'échantillonnage systématique avec classement aléatoire et PPT. Pour le renouvellement des logements et des grappes, on procède de la manière décrite au paragraphe précédent. Quelques strates urbaines font l'objet d'un échantillonnage à trois degrés. Au cours de la première étape, on choisit les UPE à l'intérieur de la strate, tandis qu'au cours de la deuxième étape, on sélectionne les grappes à l'intérieur des UPE. La dernière étape consiste, comme d'habitude, à choisir les logements à l'intérieur des grappes. Le même mode de renouvellement s'applique à chaque étape de l'échantillonnage. On peut procéder à une rotation des UPE à l'intérieur des strates et à une rotation des grappes à l'intérieur des UPE urbaines.

On procède aux remplacements des premières unités choisies et aux remplacements ultérieurs jusqu'à épuisement des unités dans l'échantillon, conformément à la règle de la durée de vie minimale. Cette règle tente de retarder le renouvellement des grappes après la sélection initiale, une mesure très importante dans les secteurs ruraux, où le renouvellement peut impliquer l'engagement de nouveaux intervieweurs ; toutefois, ce concept a été appliqué à tous les secteurs. On neutralise le biais créé par l'allongement de la durée de vie de la sélection initiale en réduisant la durée des sélections suivantes. La deuxième unité demeure dans l'échantillon jusqu'au mois de sortie normalement fixé sans correction en fonction de la durée de vie minimale. Les unités suivantes demeurent actives pendant leur durée de vie complète.

L'équation suivante doit être satisfaite pour que les probabilités de la sélection soient non biaisées :

$$K_1 + K_2 \leq R_{\min} + 1$$

où K_1 est le nombre minimum d'origines pour l'unité initialement choisie, K_2 est le nombre minimum d'origines pour les unités de remplacement suivantes, $R_{\min}$ est la plus petite FSI de toutes les unités de la strate.

On établit la valeur de K_1 selon les règles suivantes, en commençant par une valeur de base de b.

Si $b < (R_{\min}+1)/2$ alors K_1 est un nombre aléatoire dans l'intervalle $[b,\ R_{\min}-b+1]$.

Si $b \geq (R_{\min}+1)/2$ alors $K_1 = int(R_{\min}/2)+1$.

Une sélection initiale ayant une durée de vie r_1 inférieure à K_1 sera prolongée de $K_1 - r_1$. La durée de vie de la sélection suivante sera réduite de la même valeur $K_1 - r_1$. Il s'ensuit donc que $K_1 + K_2 \leq R_{\min} + 1$.

D'après notre expérience empirique au moment de la conception du plan, nous avons optimalisé la plupart des strates avec un minimum de quatre origines (deux années dans l'échantillon), tandis que la base d'appartements a été optimalisée avec deux origines. Comme la base d'appartements est ouverte, la règle de la durée de vie minimale s'applique également aux nouvelles sélections, quoique, dans ces cas, on établisse une base de quatre origines. Nous faisons ce choix surtout pour des raisons de commodité, puisque cela nous évite d'établir une nouvelle liste de grappes ayant une durée de vie très courte.

L'unité de deuxième degré demeure dans l'échantillon jusqu'à son mois de sortie normalement fixé. Les sélections suivantes demeurent dans l'échantillon pendant leur durée de vie complète.

Grandes villes, base d'appartements. Dans chaque strate de la base d'appartements, qui prend la forme d'une liste ouverte d'immeubles, on choisit les immeubles d'appartements par échantillonnage systématique avec PPT (et par échantillonnage systématique avec PPT dans les strates à faible revenu). Dans chaque immeuble d'appartements sélectionné, on prélève un échantillon systématique de cinq logements.

Autres secteurs urbains. Dans presque tous les autres secteurs urbains, la première étape consiste à choisir des grappes ou des SD au moyen de la méthode des groupes aléatoires RHC, puis à sélectionner les logements. Le nombre de logements sélectionnés par unité primaire varie, de trois (pour une grappe) à dix (pour un SD), parce que le plan couvre une grande variété de secteurs urbains et semi-urbains.

Dans certains cas spéciaux, représentant moins de 1 % de l'échantillon, la première étape d'échantillonnage consiste à choisir deux villes dans une strate à l'aide de la méthode d'échantillonnage systématique avec classement aléatoire et PPT. On choisit alors un multiple de six grappes (généralement 12 ou 18) dans chaque ville au moyen de l'échantillonnage systématique avec classement aléatoire et PPT. Enfin, un échantillon systématique de trois logements est prélevé dans chaque grappe. Soulignons qu'ici, on entend par « ville » une ville en tant que telle, deux petites villes considérées comme une seule ou une section d'une ville plus grande. On utilise cette méthode dans les cas où il n'est pas possible de confier à un intervieweur une tâche raisonnable au moyen de l'échantillonnage décrit plus haut.

Régions éloignées. La portion nordique des sept provinces non maritimes est, en grande partie, peu peuplée. Il faut donc utiliser une méthode d'échantillonnage spéciale pour ces régions. Sauf pour une exception discutée plus loin, l'échantillon est sélectionné en deux étapes. La première étape consiste à établir un échantillon d'agglomérations, que l'on appellera lieux, et de SD. En raison des grandes distances qu'il faut couvrir pour réaliser des interviews en régions éloignées, les lieux comptant moins de dix ménages ou 25 personnes sont omis du plan. De même, les SD comptant moins de 25 ménages sont également omis. Malgré ces omissions, le plan couvre environ 90 % de la population des régions éloignées de chaque province.

On prélève un échantillon des SD et des lieux à l'aide de la méthode systématique avec PPT, après que les unités ont été triées par nombre de ménages. On sélectionne ensuite un échantillon des ménages à l'aide de la méthode d'échantillonnage systématique. Si la taille d'un lieu ou d'un SD est trop grande pour en établir une liste convenable, on le répartit en grappes traitables.

Le Québec compte deux régions éloignées. Dans le cas de la première strate, on applique la méthode décrite ci-dessus, alors que dans le cas de l'autre on applique un échantillonnage à trois degrés. Cette dernière strate compte huit villes qui forment les unités primaires. Les unités secondaires sont les grappes obtenues en subdivisant les grands SD et en regroupant les petits SD. L'objectif est d'obtenir des grappes de 100 ménages, avec une tolérance d'environ 50 ménages. Deux villes sont sélectionnées au moyen de la méthode d'échantillonnage systématique avec classement aléatoire et PPT. On choisit ensuite trois grappes par ville au moyen de la même méthode, et finalement, on prélève systématiquement neuf ménages dans chaque grappe.

Renouvellement de l'échantillon

Chaque mois, une portion de l'échantillon de l'EPA est remplacée. Le renouvellement des unités d'échantillonnage est effectué à chaque étape du plan d'échantillonnage. L'unité ultime de la sélection, le logement, est remplacée tous les six mois, tandis que les unités plus globales demeurent plus longtemps dans l'échantillon. On a fixé à six mois la période de renouvellement des ménages, car cette durée représentait un compromis entre le coût du renouvellement et l'augmentation du taux de non-réponse qui pourrait se produire si l'on demandait aux ménages de participer à l'enquête plus longtemps.

Pour assurer aux intervieweurs une charge de travail uniforme et réduire au minimum le biais associé au nombre de mois pendant lesquels un ménage a été sondé, on a adopté un mode de renouvellement selon lequel un sixième de l'échantillon est remplacé chaque mois. Pour ce faire, on attribue à chaque grappe un numéro de renouvellement de 1 à 6 qui détermine quels seront les mois du renouvellement : si le numéro est 1, alors le renouvellement des ménages de la grappe a lieu en janvier et en juillet, si c'est 2, en février et en août, et ainsi de suite.

Méthode de renouvellement

Secteurs où l'on applique la méthode des groupes aléatoires. Cette méthode de sélection a été décrite à la fin du chapitre précédent. Pour la *première* grappe choisie dans chaque groupe aléatoire, on génère au hasard deux nombres entre 1 et la fraction de sondage inverse (FSI) de

Secteurs ruraux. Dans le nouveau plan, l'échantillonnage à l'intérieur des strates finales rurales se fait par la sélection des SD au premier degré d'échantillonnage suivie de la sélection des logements au deuxième degré. Les SD sont sélectionnés à l'aide d'un échantillonnage systématique avec classement aléatoire et PPT, qui est décrit au paragraphe suivant. PPT signifie probabilité proportionnelle à la taille, et ici, la taille de l'unité d'échantillonnage correspond au nombre de ménages dans l'unité au cours du recensement de 1991. À l'intérieur des SD sélectionnés, on choisit au hasard un échantillon systématique de logements. Habituellement, on choisit de cette manière 10 logements par SD. Dans l'ancien plan, on utilisait trois degrés d'échantillonnage dans la plupart des secteurs ruraux, les unités primaires d'échantillonnage (UPE) formées de groupes de SD étant sélectionnées en premier lieu. Ensuite, on sélectionnait les SD à l'intérieur des UPE, puis les logements à l'intérieur des SD. Cette approche en trois étapes était particulièrement utile lorsque la plupart des interviews étaient réalisées sur place, puisque l'UPE correspondait en gros à une affectation d'intervieweur et les déplacements en étaient facilités. Comme cinq sixièmes des interviews sont maintenant réalisées par téléphone, on peut appliquer un plan plus simple dans la plupart des secteurs du pays, et l'étape des UPE a été éliminée, sauf dans certaines régions éloignées. On trouvera dans Mantel et coll. (1994) une description des diverses stratégies d'échantillonnage pour les secteurs ruraux qui ont été comparées dans le cadre de l'actuel projet de remaniement.

La méthode d'échantillonnage systématique avec classement aléatoire et PPT élaborée par Hartley et Rao (1962) est utilisée dans le cadre de l'EPA depuis les années 1970. Selon cette méthode, les unités primaires d'une strate sont classées aléatoirement et un échantillon systématique avec PPT de la taille souhaitée est ensuite sélectionné. Dans le nouveau plan, l'unité primaire est le SD, six SD étant sélectionnés par state. On trouvera dans Cochran (1977) une description de l'échantillonnage systématique avec PPT.

Dans les secteurs ruraux comptant une faible densité de peuplement, on a utilisé un autre plan d'échantillonnage. On a formé des unités primaires géographiquement compactes constituées de six SD, et deux ou trois de ces unités ont été sélectionnées à l'aide de la méthode d'échantillonnage systématique avec classement aléatoire et PPT. Dans les unités primaires sélectionnées, on a établi un échantillon systématique de logements dans chacun des six SD, ce qui veut dire que l'on ne fait pas de sous-échantillonnage à l'intérieur des SD. On a adopté ce plan avec grappes pour qu'il y ait assez de travail pour occuper un intervieweur dans les secteurs peu peuplés se trouvant dans l'échantillon.

Grandes villes, base des non-appartements. Dans les secteurs urbains, la première étape consiste à sélectionner des grappes. Comme le fichier du réseau routier ne couvre pas parfaitement les villes, en particulier leurs quartiers périphériques, il est impossible de former des grappes partout, et on utilise alors comme unités primaires des SD et des sections de SD. À la deuxième étape, on choisit à l'intérieur de la grappe un échantillon systématique de logements. Pour la base d'appartements, l'immeuble d'appartements tient lieu de grappe. On peut donc dire que tant dans les secteurs ruraux qu'urbains, la norme est maintenant un échantillonnage à deux degrés.

Dans les secteurs urbains, on procède à la sélection des grappes et des SD au moyen de la méthode des groupes aléatoires conçue par Rao, Hartley et Cochran (1962) ; voir aussi Cochran (1977). On a introduit cette méthode dans les années 1970 parce qu'elle permet de faire une révision relativement directe des probabilités de sélection des grappes. De telles révisions peuvent être nécessaires dans les secteurs urbains ayant connu une importante croissance démographique. La souplesse de cette méthode permet également d'apporter les ajustements nécessaires lorsqu'il faut modifier la taille de l'échantillon de l'EPA, ce qui arrive à l'occasion. Nous présentons dans les lignes qui suivent un aperçu de l'application à l'EPA de la méthode des groupes aléatoires de Rao-Hartley-Cochran (RHC). On trouvera de plus amples renseignements dans Singh et coll. (1990).

Pour une strate donnée, dans laquelle la méthode RHC est appliquée, les grappes sont affectées aléatoirement à six groupes appelés groupes aléatoires. Dans la mesure du possible, on met dans chaque groupe un nombre identique de grappes, la variation étant, au plus, d'une grappe. Dans certains groupes on utilise un multiple de six groupes. On choisit une grappe dans chaque groupe aléatoire selon une probabilité proportionnelle à sa taille. Ainsi, si une grappe est deux fois plus grande qu'une autre, elle aura deux fois plus de chances d'être choisie que la seconde.

À l'intérieur des grappes urbaines sélectionnées dans les strates non identifiées à revenu élevé, on sélectionne un échantillon systématique de logements. À Montréal, à Toronto et à Vancouver, on sélectionne six logements par grappe. Dans les autres secteurs urbains, on choisit huit logements par grappe. Dans les grappes des strates à revenu élevé, on prélève un échantillon de quatre logements.

procédé ainsi pour corriger des déséquilibres résultant de réductions de l'échantillon effectuées sur l'ancien plan (les provinces comptant des RMR plus grandes ont été davantage touchées par les réductions parce que celles-ci avaient tendance à être concentrées dans les grandes RMR).

Répartition de l'échantillon de base entre les RE à l'intérieur des provinces. Dans chaque province, on a réparti l'échantillon de base entre les RE en proportion de la taille de ces régions, cette taille étant mesurée selon le nombre de logements privés occupés dans la RE d'après les données du recensement de 1991. Cette répartition visait principalement à optimaliser les données à l'échelle provinciale. Toutefois, comme les RE peu peuplées reçoivent un échantillon trop petit si l'on suit une répartition proportionnelle stricte, on a établi une taille minimale, soit 200 ménages par RE. En Alberta, où le minimum a été établi à 300 ménages, on a sacrifié quelque peu l'efficacité au niveau provincial pour avantager les RE de petite taille.

Répartition à l'intérieur des RE. Les échantillons répartis à l'étape précédente ont ensuite été répartis proportionnellement à l'intérieur de chaque RE entre les intersections RE-REAE. Encore une fois, il s'agit là d'une répartition presque optimale de l'échantillon.

Répartition de l'échantillon financé par DRHC entre les REAE. La répartition de l'échantillon de base se terminait avec l'étape précédente. L'étape suivante consistait à répartir l'échantillon de DRHC de 16 500 ménages. Comme le bailleur de fonds de cet échantillon est DRHC, qui veut s'assurer que les estimations sur la population active des REAE sont de qualité suffisante, l'échantillon a été réparti entre les REAE pour optimaliser l'amélioration du CV des *chômeurs*, en ciblant l'échantillon vers les régions qui avaient un CV élevé basé sur l'échantillon de base. En répartissant l'échantillon de cette manière, on a pu obtenir un CV de 10 % ou moins pour les données sur les *chômeurs* dans chaque REAE. On a établi pour chaque REAE une taille d'échantillon minimale de 600 ménages.

Le tableau A2 de l'annexe A présente les répartitions de base et totales par province pour l'ancien et le nouveau plan au moment du remaniement, de même que la répartition effectuée après la réduction de l'échantillon réalisée en 1995.

Détermination des CV. À différentes étapes du processus de répartition, on a calculé les CV de la caractéristique *chômeurs*. La variance de cette caractéristique est une fonction du taux de chômage. On a utilisé les taux de chômage moyens des provinces et des régions infra-provinciales tirés de l'EPA pour la période de 1984 à 1992. On a choisi cette période pour obtenir des taux typiques de ceux que l'on risque de connaître au cours de la durée de vie du plan d'échantillonnage. En outre, puisque le plan de l'EPA fait appel au groupement, on peut exprimer la variance comme étant une fonction d'un effet du plan d'échantillonnage aussi bien que du taux de chômage. (On entend par effet du plan d'échantillonnage d'un estimateur le ratio de la variance observée selon le plan retenu à la variance prévue pour un échantillon aléatoire simple de la même taille.) Les effets du plan utilisés étaient fondés sur les effets estimatifs du plan de 1989 à 1992. On a calculé des moyennes lissées pour chaque RE. On trouvera dans Mian et Laniel (1994) une description plus détaillée de la façon dont les CV ont été calculés.

Réduction de la taille de l'échantillon. Après la mise en place du nouveau plan, on a réduit la taille de l'échantillon de base de 6 500 ménages. Cette réduction est entrée en vigueur en juillet 1995. Grâce à l'efficacité améliorée du plan, les CV des estimations nationales et provinciales sont demeurés, après la réduction, identiques à ce qu'ils étaient avant le remaniement avec un échantillon plus grand. L'échantillon a été diminué du même pourcentage dans chaque province. La taille des échantillons actuels de chaque RE et REAE est présentée au tableau A3 de l'annexe A.

Sélection de l'échantillon

Degrés d'échantillonnage. L'un des principaux changements apportés dans le nouveau plan de l'EPA est l'utilisation d'un échantillonnage à deux degrés seulement dans presque toutes les régions. Le premier degré est un échantillon aréolaire. Pour choisir les unités primaires, on prépare (et tient à jour) une liste des logements sur le terrain. Le deuxième degré consiste alors à sélectionner dans chaque liste un échantillon de logements.

Le remplacement de l'échantillonnage à trois degrés de l'ancien plan par un échantillonnage à deux degrés comporte plusieurs avantages dans les secteurs ruraux. Non seulement la méthode est-elle plus simple, mais l'échantillonnage à deux degrés est statistiquement plus efficace que l'ancien. Comme l'emplacement des unités primaires fait l'objet de moins de contraintes, la dispersion de l'échantillon est améliorée. Cela comporte un avantage pour les estimations des secteurs de petite taille ; voir Singh et coll. (1994).

CHAPITRE 3 - Répartition, sélection et renouvellement de l'échantillon

Le remaniement n'a pas eu pour effet de modifier la taille totale de l'échantillon mensuel de l'EPA, qui est demeuré identique à ce qu'elle était avec l'ancien plan, c'est-à-dire 58 850 ménages. Toutefois, dans le cadre du remaniement, l'échantillon a été réparti afin de mieux répondre au besoin de données de qualité à différents niveaux géographiques. On parle ici des échelles nationale et provinciale, des RMR, et pour la première fois, des régions d'assurance-chômage, maintenant renommées régions économiques d'assurance-emploi. Voici les objectifs de fiabilité qui étaient visés.

Canada et les provinces - Maintenir ou améliorer le coefficient de variation (CV) pour les *chômeurs*, par rapport à l'ancien plan, c'est-à-dire environ 2 % pour le Canada et de 4 à 7 % pour les provinces.

REAE/RMR - CV de 15 % ou moins pour les données trimestrielles sur les *chômeurs*. Afin d'assurer un CV de ce niveau, on a établi à 600 ménages la taille minimale de l'échantillon par REAE.

Bien qu'il ne s'agisse pas d'une exigence officielle, on a établi à 25 % ou moins le CV relatif aux estimations trimestrielles des RE concernant les *chômeurs*, bien qu'il ait fallu faire quelques regroupements. Il y a 72 RE, mais pour les fins de la répartition, ce nombre a été réduit à 68 par suite d'un regroupement de deux RE au Québec, au Manitoba, en Saskatchewan et en Colombie-Britannique. L'objectif de CV se rapporte alors aux régions regroupées. Le tableau 3 présente le nombre de régions infra-provinciales dans chaque province au moment du remaniement.

Tableau 3. RE, REAE et RMR par province

Province	RE	REAE	RMR
Terre-Neuve	4	3	1
Î.-P.É.	1	1	0
Nouvelle-Écosse	5	5	1
Nouveau-Brunswick	5	4	1
Québec	16	13	6*
Ontario	11	18	10*
Manitoba	8	3	1
Saskatchewan	6	4	2
Alberta	8	4	2
C.-Britannique	8	6	2
Canada	72	61	25*

*La RMR d'Ottawa-Hull est comptée à la fois en Ontario et au Québec.

Comme nous l'avons noté dans la section sur la stratification, les intersections RE-REAE sont devenues des strates. Par conséquent, elles ont également servi de secteurs de base pour les fins de la répartition de l'échantillon. Comme tant pour les RE que pour les REAE, on utilise les divisions de recensement comme composante de base, il y a seulement 133 intersections d'un bout à l'autre du Canada.

La répartition de l'échantillon dans les provinces et dans les régions interprovinciales a été discutée avec les représentants des provinces et les principaux utilisateurs des données de l'EPA. La répartition finale a également subi certaines contraintes opérationnelles. On a étudié plusieurs stratégies de répartition, notamment la répartition de Neyman, celle de Kish, la répartition proportionnelle, la méthode puissance et la répartition selon la racine carrée. Mian et Laniel (1994) font le résumé de ces diverses stratégies. Dans ces lignes, nous nous contentons de décrire la démarche que nous avons appliquée.

Au moment du remaniement, l'échantillon total de l'EPA comptait un échantillon de base de 42 310 ménages et un échantillon financé par DRHC de 16 500 ménages. La répartition de l'échantillon de base n'était pas optimale pour les estimations provinciales et infraprovinciales étant donné les changements survenus dans la taille d'échantillon et la population depuis le remaniement précédent. La stratégie globale consistait à répartir d'abord l'échantillon de base afin d'optimaliser les estimations provinciales et nationales. On a ensuite procédé à la répartition de l'échantillon de DRHC afin de compléter l'échantillon de base dans les régions d'assurance-emploi qui en avaient le plus besoin (généralement, des secteurs dont la population est relativement faible).

Répartition de l'échantillon de base entre les provinces. Outre les exceptions suivantes, nous avons conservé la même taille pour les échantillons provinciaux de base (soit celle existante avant le financement de DRHC). Les exceptions sont : un transfert d'échantillon de la Saskatchewan au Manitoba et de l'Alberta à la Colombie-Britannique. La portion d'échantillon transférée était juste suffisante pour donner à chacune des deux provinces impliquées dans le transfert le même CV relativement aux *chômeurs*, étant donné l'échantillon de base. On a

taille de l'échantillon était d'au moins 50 logements), on a utilisé cette ancienne stratification dans le nouveau plan. Dans certains cas, des ajouts au secteur urbain ont été affectés manuellement à une strate. On utilise ici le terme *plan FV* parce que dans le cadre des remaniements antérieurs, les dénombrements tirés de feuilles de visites du recensement étaient utilisés pour former les grappes.

Groupement en strates finales. Pour réduire les coûts sur le terrain, on ne sélectionne pas directement les ménages qui feront partie de la strate finale. Chaque strate est plutôt divisée en grappes, un échantillon de grappes étant ensuite choisi à l'intérieur de la strate. Ensuite, un échantillon de ménages est choisi dans chaque grappe sélectionnée. On trouvera dans la section traitant de la sélection de l'échantillon une description de la méthode de sélection des grappes.

Dans les secteurs ruraux, les SD servent généralement de grappes. Dans les secteurs urbains, on utilise une variété de grappes. Dans les petites villes, où l'on a conservé la stratification du plan précédent, on a également utilisé les anciennes grappes. Celles-ci ont été formées manuellement à l'aide des FV du recensement : on a combiné des côtés d'îlot jusqu'à ce que la grappe soit de la taille souhaitée. Le dénombrement de la population de ces grappes a été mis à jour au moyen des données du recensement de 1991 et, dans certains cas, on a dû modifier les anciennes grappes parce que des changements importants étaient survenus depuis l'ancien plan. Dans les secteurs urbains, où la stratification repose sur les SD, on a également utilisé ces dernières unités pour en faire des grappes. Dans certains cas, les grands SD ont été divisés en deux grappes.

Dans les secteurs urbains de grande taille, c'est-à-dire ceux qui font partie du FRR, on a utilisé une nouvelle méthode de groupement automatisée, qui a remplacé l'ancien travail manuel à forte intensité de main-d'oeuvre. Pour le nouveau plan, le PPCAO, qui avait été utilisé pour former les SD du recensement de 1991, a été modifié par le personnel de la Division de la géographie de Statistique Canada pour former les grappes de l'EPA. Le programme combine les côtés d'îlot pour produire des grappes de 150 à 200 logements (200 à 250 logements à Montréal, Toronto et Vancouver) en moyenne.

Les nouvelles grappes urbaines ont environ trois fois la taille des grappes utilisées dans le plan précédent. Cette augmentation contribuera à atténuer le problème attribuable à la croissance démographique accélérée qui se produit occasionnellement dans les secteurs urbains puisque l'incidence relative de la croissance aura tendance à être moindre dans les grappes de grande taille. La grande taille des grappes réduit également la fréquence de leur renouvellement.

Le tableau 2 donne un aperçu des types d'unités primaires utilisées pour l'ensemble de l'échantillon de l'EPA. La taille, qui se rapporte au nombre de ménages dans une unité typique, peut varier considérablement à l'intérieur d'un type d'unité donné. Le rendement est le nombre de ménages sélectionnés dans le cadre de l'EPA pour être sondés au cours d'un mois donné.

On trouvera à l'annexe D une liste de tous les secteurs urbains du Canada constituant les strates ou les groupes de strates de l'EPA, c'est-à-dire tous les secteurs urbains dans lesquels se trouve toujours un échantillon de l'EPA. Cela correspond à l'ancien concept d'unité autoreprésentative dans les plans antérieurs. Tout secteur non énuméré à l'annexe D n'est donc pas une strate. Il s'agit donc plutôt d'une unité primaire d'échantillonnage ou d'une partie d'une strate urbaine plus grande ou d'une strate rurale. L'annexe D présente également les types de grappes que l'on trouve dans chaque secteur.

Tableau 2. Principales unités primaires, tailles et rendements

Secteurs	Unité d'échantillonnage	Taille (nombre de ménages par unité)	Rendement (nombre de ménages dans l'échantillon)
Toronto, Montréal, Vancouver	grappe	200-250	6
Autres villes	grappe	150-200	8
Base d'appartements	immeubles d'appartements	variable	5
Majorité des secteurs ruraux et secteurs urbains non-FRR	SD	300	10

doit contenir au moins 30 logements et le revenu moyen de l'ensemble de la base doit être d'environ 15 000 $. On a créé des bases de faible revenu dans sept villes : Montréal, Ottawa (excluant Hull), Toronto, Winnipeg, Calgary, Edmonton et Vancouver.

Cas spéciaux - À Calgary et à Edmonton, on a ajouté à la base des immeubles dont le revenu moyen était supérieur à 20 000 $ afin que le rendement de l'échantillon atteigne le niveau souhaité. À Montréal, la base de faible revenu est entièrement comprise dans la RE 40.

À la différence du reste de la base d'appartements, la base de faible revenu n'est pas ouverte, puisque l'on ne sait pas quel sera le revenu moyen des résidents d'un nouvel immeuble d'appartements.

Parmi les sept grandes villes ayant une base de faible revenu, seule Toronto compte assez de logements pour que la base fasse l'objet d'une stratification. La SDR de Toronto et le reste de la RMR de Toronto ont toutes les deux une base d'appartements à faible revenu.

En ce qui concerne les appartements qui ne sont pas à faible revenu, on a tenté de créer des listes à l'intérieur de superstrates géographiques. On a pu le faire à Halifax (2 strates d'appartements), à Québec (2), à Montréal (4), à Ottawa (3), à Toronto (6), à Hamilton (2), à Kitchener (2) et à Vancouver (4). Cela produit des ventilations géographiques des strates d'appartements globales à l'intérieur de ces grandes villes.

Enfin, on a également tenté de subdiviser les strates d'appartements selon la taille des immeubles. Dans chaque strate, on a classifié les immeubles selon la taille : moins de 100 logements, entre 100 et 199 logements et 200 logements et plus. S'il y avait assez d'appartements dans une catégorie de taille pour produire un échantillon de 30 logements, alors on en a fait une strate distincte. Sinon, on l'a regroupée avec une autre catégorie.

La stratification de la base d'appartements est résumée au tableau 1.

Tableau 1. Strates d'appartements

RMR	Strates géographiques	Nombre total de strates
Halifax	2	2
Québec	2	2
Montréal*	4	9
Ottawa -Hull*	3	6
Oshawa	1	2
Toronto*	6	16
Hamilton	2	4
St-Catharines	1	1
Kitchener	2	2
London	1	2
Windsor	1	2
Winnipeg*	1	6
Saskatoon	1	1
Calgary*	1	3
Edmonton*	1	3
Vancouver*	4	6
Victoria	1	1
TOTAL	34	68

Note : i) un astérisque (*) dénote que cette grande ville compte au moins une strate de faible revenu, ii) le nombre total de strates comprend les strates à faible revenu.

3.1 Petites villes - Plan d'échantillonnage des SD (voir annexe D). Dans toutes les villes, sauf les plus petites, on a formé des strates finales optimales non compactes et non contiguës en utilisant les SD comme unités de stratification. À Sydney (Nouvelle-Écosse), on a d'abord créé des strates compactes et contiguës pour ensuite créer les strates finales à l'intérieur de chaque superstrate.

3.2 Petites villes : plan d'échantillonnage FV (Feuille de visites) (voir annexe D). Dans les secteurs urbains de petite taille classés « autoreprésentatifs » dans l'ancien plan (il constituait au moins une strate urbaine dont la

Lorsque plus d'un niveau de stratification existe dans une grande ville, nous ferons référence au plus bas niveau, à la plus petite strate en tant que strate finale. On a conçu la strate finale pour qu'elle ait une taille attendue d'au moins 48 ménages (35 ménages à Toronto, Montréal et Vancouver). La taille prévue des échantillons est plus faible dans ces trois grandes villes parce que, dans le nouveau plan, le rendement de l'échantillon par grappe a été choisi pour être environ le double du rendement de l'ancien plan, ce dernier ayant été quelque peu inférieur dans ces trois villes.

2.1 Grandes villes - secteurs FRR (ou PPCAO). Ces secteurs urbains font partie du fichier du réseau routier de la Division de la géographie. Ils comprennent les 25 RMR et 20 des plus grandes agglomérations de recensement (AR). Dans ces secteurs, il peut y avoir jusqu'à trois niveaux de stratification. À l'intérieur d'une RMR ou d'une AR, si la taille prévue de l'échantillon d'une municipalité (c.-à-d. une subdivision de recensement ou SDR) est d'au moins 240 ménages, (180 à Toronto, à Montréal ou à Vancouver), alors la municipalité comme telle devient une strate. Si la municipalité est assez importante pour former plus de cinq strates finales, elle fait l'objet d'une stratification optimale en groupes appelés superstrate, dont le rendement sera de trois (parfois quatre ou cinq) strates finales. On forme les superstrates en utilisant les secteurs de recensement (SR) comme unités de stratification (ou les SDR dans les quartiers périphériques non dépouillés des villes). Ces superstrates sont géographiquement compactes et contiguës. À Toronto, Montréal et Vancouver, l'objectif était de créer des superstrates dont le rendement serait de six strates finales plutôt que de trois.

Si la taille d'une SDR n'est pas suffisante pour former plus de cinq strates finales, on la regroupe avec d'autres SDR analogues. On traite ensuite ce groupement de la manière décrite au paragraphe précédent sur les superstrates (c.-à-d. que s'il est assez grand, on y forme une strate optimale, etc.).

Les superstrates sont divisées de façon optimale en strates finales ; elles sont non compactes et non contiguës et leur rendement est de 48 ménages, sauf à Toronto, à Montréal et à Vancouver où il est de 36 ménages.

2.2 Grandes villes - strates à revenu élevé. Pour la première fois, on a formé des strates à revenu élevé dans les neuf grandes villes où cela était possible, soit Montréal, Ottawa, Toronto, Hamilton, London, Winnipeg, Edmonton, Calgary et Vancouver. Dans chacune de ces grandes villes, 3 % des secteurs de dénombrement (SD) affichant les revenus moyens des ménages les plus élevés selon le recensement de 1991 forment la strate à revenu élevé. Chaque strate doit compter au moins 24 ménages. Dans cinq villes, il y avait assez de SD pour constituer deux strates à revenu élevé ou plus. Le tableau A1 de l'annexe A donne un aperçu de cette opération. Les villes qui ne sont pas dans cette liste ne comptaient pas assez de SD avec un revenu moyen des ménages élevé (environ 100 000 $) pour former une strate séparée. On trouvera dans Chen et coll. (1994) de plus amples détails sur les strates à revenu élevé.

On pense que l'introduction des strates à revenu élevé devrait, avec le temps, stabiliser la représentation dans l'échantillon des ménages à revenu élevé. Cette mesure profitera à certaines enquêtes, notamment l'Enquête sur les finances des consommateurs, qui se sert du plan ou de l'échantillon de l'EPA et qui recueille des données relatives au revenu. Cela facilitera également la collecte de données sur les gains dans le nouveau questionnaire de l'EPA. De plus, il pourrait être plus facile d'établir si la propension de non-réponse est plus élevée pour les ménages à revenu élevé. Si cette hypothèse est confirmée, on pourra prendre des mesures à cet égard.

2.3 Grandes villes - la base d'appartements. Dans le cadre de l'EPA, on tient à jour une liste d'immeubles d'appartements dans les grandes RMR depuis les années 1960. À l'heure actuelle, cette liste est utilisée comme base d'échantillonnage dans 18 grandes villes : Halifax, Québec, Montréal, Hull, Ottawa, Oshawa, Toronto, Hamilton, St-Catharines, Kitchener, London, Windsor, Winnipeg, Saskatoon, Calgary, Edmonton, Vancouver et Victoria. Dans le cadre de l'EPA, on entend par immeuble d'appartements tout immeuble comptant cinq étages habités et d'au moins 30 logements. Dans chaque ville, tout immeuble neuf est ajouté au bas de la liste de cette ville. Comme l'échantillonnage des appartements est systématique, il est probable que tout nouvel immeuble d'appartements entre dans l'échantillon dès que sa construction est achevée.

Une nouvelle caractéristique du plan de l'EPA est la formation d'une base d'immeubles d'appartements à faible revenu. Contrairement à la strate à revenu élevé, on a trouvé plus pratique d'utiliser les immeubles d'appartements plutôt que les SD pour les fins de la stratification des ménages à faible revenu.

On ajoute un immeuble d'appartements à la base de faible revenu si le revenu moyen des ménages qui y habitent est inférieur à 20 000 $ selon le recensement de 1991. Pour qu'une base de faible revenu existe dans une ville, elle

dans Drew et coll. (1985) et dans Singh et coll. (1990). On appelle *stratification optimale* la stratification fondée sur cet algorithme.

Variables de stratification. Les variables utilisées dans le programme de stratification, présentées ci-dessous, ont toutes été utilisées dans le plan antérieur de l'EPA. Cependant, la ventilation sectorielle des emplois est plus détaillée, en particulier dans les secteurs manufacturiers et des services. La seule stratification entièrement nouvelle est fondée sur la langue maternelle. Pour chaque unité de stratification, on a codé trois variables linguistiques : le nombre de personnes qui déclarent avoir comme langue maternelle l'anglais, le français ou une autre langue. Enfin, la variable revenu a reçu trois fois la pondération des autres variables de stratification.

On a utilisé les données du recensement de 1991 aux fins de la stratification. Voici les variables qui ont été utilisées :

Nombre de personnes occupées dans les secteurs suivants:
- agriculture
- foresterie ou pêcheries
- mines
- secteur manufacturier - biens de consommation
- secteur manufacturier - caoutchouc, plastique, cuir
- secteur manufacturier - textiles et vêtements
- secteur manufacturier - meubles, pâtes et papier, imprimerie, bois
- secteur manufacturier - métaux et minéraux
- secteur manufacturier - pétrochimie et chimie
- construction
- transports
- services - commerciaux
- services - financiers
- services - personnels / entreprises
- services - gouvernement

Nombre total de personnes occupées
Revenu total
population de 15 ans et plus
population de 15 à 24 ans
population de 55 ans et plus
nombre de ménages d'une personne
nombre de ménages de deux personnes
nombre de logements en propriété
loyer brut total
population ayant fait des études secondaires
langue maternelle anglaise
langue maternelle française
langue maternelle autre que l'anglais ou le français

Le choix des variables de stratification a été adapté à chaque région faisant l'objet d'une stratification optimale. À l'intérieur des régions faisant l'objet d'une stratification optimale, on a obtenu les variables ci-dessus à partir des données du recensement de 1991. Si une variable représentait moins de 2 % de la population totale, alors elle était écartée. Pour les catégories comme les services, si une sous-catégorie, telle les services financiers, comptait trop peu d'effectif, alors on a plutôt utilisé la variable globale. Une catégorie est considérée significative si elle représente plus de 2 % de la population.

Types de secteurs aux fins de la stratification. On peut diviser le plan de l'EPA en trois genres de secteurs : 1) les secteurs ruraux, 2) les grandes villes et 3) les petites villes. Pour les besoins de la stratification, chacun de ces secteurs peut ensuite être subdivisé, comme on l'indique dans les lignes qui suivent.

1. Secteurs ruraux. Dans les secteurs ruraux, les strates sont généralement formées en regroupant manuellement deux ou trois divisions de recensement à l'intérieur d'une intersection RE-REAE. Les décisions relatives à la formation de ces strates géographiques ont été faites en conjonction avec les décisions touchant les unités primaires d'échantillonnage les plus pertinentes (un échantillonnage à deux degrés des secteurs de dénombrement par rapport à un échantillonnage à trois degrés des unités primaires d'échantillonnage) et la pertinence de former des strates urbaines et rurales séparées. On a procédé à une stratification optimale à l'intérieur de la strate géographique dans tous les cas où la taille de la population était suffisante. Règle générale, les strates rurales du nouveau plan sont plus petites que les strates rurales du plan précédent.

2. Grandes villes (population de 50 000 ou plus) : Dans 17 RMR, il y avait un nombre suffisant d'immeuble d'appartements pour former une base séparée, appelée base d'appartements, qui est décrite au point 2.3 ci-dessous. Outre la base d'appartements, le reste de chaque centre urbain comporte une base aréolaire. De plus, lorsque c'était réalisable, on a formé une strate distincte avec les secteurs à revenu élevé (voir le point 2.2). Le reste des logements forme la strate ordinaire, qui est décrite ci-dessous dans la section *secteur FRR* (fichier du réseau routier). Comme on le notera plus loin (2.1) le concept de secteur FRR correspond aux secteurs établis au moyen d'un Programme de partage par circonscription assisté par ordinateur (PPCAO).

CHAPITRE 2 - Stratification et formation des unités d'échantillonnage

La population du Canada habite différentes zones géographiques, comme les provinces et les régions, lesquelles reposent sur des définitions standard. Habituellement, les échantillonneurs subdivisent ces régions en strates, à partir desquelles ils sélectionnent séparément les échantillons. Si la population d'une strate donnée est relativement homogène, alors la taille de l'échantillon nécessaire pour obtenir des estimations d'une certaine précision sera beaucoup plus petite que dans le cas d'un plan non stratifié. La stratification offre d'autres avantages : on peut utiliser des plans d'échantillonnage et des méthodes d'estimation qui diffèrent d'une strate à une autre, on peut modifier le plan seulement dans les strates qui ont connu une évolution rapide, et les strates peuvent constituer des unités opérationnelles utiles. Dans la présente section, nous décrivons la stratification utilisée dans le cadre de l'EPA.

La plupart des enquêtes utilisent deux types de strates : i) les unités géographiques standard telles que les régions métropolitaines et ii) les strates formées en combinant, selon un critère objectif, des unités plus petites comme par exemple les secteurs de dénombrement du recensement. Nous décrivons pour commencer les unités géographiques standard utilisées dans le cadre de l'EPA.

Toutes les provinces, à l'exception de l'Île-du-Prince-Édouard, sont divisées en régions économiques (RE). L'EPA utilise la région économique comme strate de base depuis les années 1960. Au cours des premières étapes du remaniement actuel, les délimitations des RE ont été révisées en consultation avec les provinces. À l'heure actuelle, on dénombre 72 RE au Canada.

Dans les remaniements précédents, les RE étaient les seules subdivisions provinciales dont on tenait compte au moment de concevoir l'enquête. En 1989, Développement des ressources humaines Canada (DRHC) a commencé à financer une augmentation de l'échantillon de l'EPA de 16 500 ménages à tous les mois. Cet échantillon est utilisé afin d'obtenir de meilleures estimations sur la population active dans les anciennes 61 (maintenant 53) régions économiques d'assurance-emploi (REAE). Voilà pourquoi le nouveau plan tient compte, aux fins de la stratification, des RE et des REAE. Pour les besoins de la répartition, on a porté moins d'attention à améliorer les estimations des RE, puisque l'échantillon de base était réparti principalement dans l'optique de produire des données provinciales et nationales, tandis que l'échantillon supplémentaire financé par DRHC ciblait les REAE.

Les deux découpages, qui comptent un nombre sensiblement égale de régions, sont définis à des fins différentes et ne coïncident généralement pas. Afin de pouvoir utiliser simultanément les deux découpages dans le cadre du nouveau plan de l'EPA, on a utilisé comme strate de base les intersections des régions. Compte tenu des chevauchements entre les REAE et les RE, on a relevé 133 intersections.

L'EPA utilise également, dans le plan actuel comme dans les plans antérieurs, un troisième découpage : les RMR. Les RMR sont des secteurs urbains comptant au moins 100 000 habitants selon le plus récent recensement. Toutes les RMR sont également des REAE.

Antérieurement, lorsqu'un nouveau plan de sondage était introduit, les définitions des RMR qu'on utilisait dataient déjà de quatre ans. Par exemple, le plan précédent est devenu parfaitement opérationnel en mars 1985 avec des définitions pour les RMR datant de juin 1981. Pour le nouveau plan, on a accéléré le travail de définition des RMR du recensement de 1996, pour les besoins du remaniement de l'EPA. Les RMR officiels de 1996 différeront de celles utilisées initialement par l'EPA, puisqu'il a fallu tenir compte des modifications formelles apportées aux limites municipales. Ces différences sont mineures et on apportera des corrections à l'EPA en fonction des définitions finales des RMR. Les RMR de 1996 servent également de REAE.

À l'intérieur des strates géographiques plus grandes, on a formé des strates plus détaillées sans égard aux contraintes géographiques. Pour ce faire, on a utilisé la même méthode que dans les plans antérieurs, c'est-à-dire un algorithme de groupement élaboré par Friedman et Rubin (1967) et modifié par Drew et coll. (1985) pour les besoins de l'EPA. L'algorithme a pour objet de répartir les unités en strates aussi homogènes que possible selon certaines variables en réduisant la somme des carrés pondérés à l'intérieur de chaque groupe. On calcule la somme des carrés pour rendre compte de l'échantillonnage des unités avec probabilité proportionnelle à la taille. On peut, au besoin, affecter différentes pondérations à différentes variables. On trouvera de plus amples renseignements sur l'algorithme

Objectifs du remaniement de l'échantillon

Le programme de remaniement actuel de l'échantillon a clôturé par la mise en place d'un nouvel échantillon à la fin de 1994. Dans le cadre de ce programme, on a procédé à des consultations approfondies visant à réévaluer non seulement la fonction principale de l'enquête, c'est-à-dire la production de données actuelles sur le marché du travail, mais aussi l'utilisation qui en est faite à Statistique Canada comme instrument central pour la réalisation des enquêtes auprès des ménages.

Un remaniement permet de mettre à jour la base de sondage, la stratification et la répartition de l'échantillon en fonction des changements survenus relativement à la taille et à la distribution de la population. Le plan en place jusqu'à la fin de 1994 utilisait les délimitations géographiques du recensement de 1981 et les dénombrements correspondants pour choisir l'échantillon et pour dériver les estimations pondérées nécessaires. Comme de nombreuses unités géographiques standard changent à chaque recensement, chaque remaniement permet d'adopter les plus récentes définitions de ces unités pour l'EPA.

Les remaniements antérieurs visaient à assouplir le cadre, l'échantillon et les systèmes pour le bénéfice d'autres enquêtes, puisque, à Statistique Canada, beaucoup d'enquêtes auprès des ménages utilisent ces éléments pour répondre à leurs propres besoins. Cet objectif est également important pour l'actuel remaniement. Un autre objectif partagé par les deux derniers remaniements était de tirer parti de l'évolution des technologies et des tâches sur le terrain pour en simplifier le plan.

Note concernant les régions d'assurance-chômage et les régions économiques d'assurance-emploi. Tel que mentionné ci-haut dans l'aperçu général de l'enquête, en 1995, les régions économiques d'assurance-emploi (REAE) remplaçaient les régions d'assurance-chômage (RAC). Pour faciliter la lecture de ce document, le nouveau nom (REAE) sera utilisé même si lors du remaniement de l'échantillon on a utilisé les régions d'assurance-chômage. On réferera à l'échantillon AE pour désigner l'échantillon supplémentaire de 16 500 ménages introduit en 1989 pour améliorer les estimations des REAE.

On fera référence à l'échantillon de base pour désigner l'échantillon restant qui comprend actuellement 35 850 ménages.

Un nouvel objectif de ce remaniement était d'utiliser l'échantillon de base pour répondre aux exigences des estimations nationales et provinciales tout en utilisant l'échantillon AE pour améliorer les estimations de REAE. Le chapitre 2 traitera de l'approche utilisée pour atteindre cet objectif.

Objet de la publication

La présente publication se veut un ouvrage de référence sur la méthodologie de l'EPA. Sont traités en détail: le plan de sondage, la méthodologie d'estimation et la qualité des données, tandis que des renvois additionnels sont donnés lorsque pertinent. Un document distinct intitulé *Guide de l'Enquête sur la population active* (disponible sur internet à www.statcan.ca/francais-/concepts/labour/index_f.htm) sert de complément au présent rapport en mettant l'accent sur les concepts et les définitions et sur les données produites dans le cadre de l'EPA.

Le présent document comprend cinq autres chapitres. Le chapitre 2 décrit la stratification et la formation des unités d'échantillonnage comme par exemple les grappes. Le chapitre 3 traite en détail de la répartition de l'échantillon, de la sélection de l'échantillon dans les secteurs urbains et ruraux et du renouvellement de l'échantillon dans le temps. Au chapitre 4, on aborde brièvement l'utilisation de l'échantillon et de la base de sondage de l'EPA par d'autres enquêtes auprès des ménages.

Le chapitre 5 donne une description détaillée du système de pondération et d'estimation de l'EPA, y compris le traitement de la non-réponse. Enfin, l'EPA comporte un programme poussé de contrôle de la qualité des données, lequel est décrit au chapitre 6.

Note : Une liste des abréviations utilisées dans ce document est présentée à l'annexe B. Un diagramme illustrant le nouveau plan de l'Enquête sur la population active est donné à l'annexe C.

sujets comme les heures de travail, les jeunes sur le marché du travail et les salaires seront périodiquement réétudiés.

L'EPA est également à la source du CD-ROM *Revue chronologique de la population active* (numéro 71F0004XCB au catalogue), qui contient des données détaillées sous forme de séries transversales et chronologiques de 1976 à l'année en cours.

L'enquête peut produire beaucoup plus d'information que ce qui est publié périodiquement. Des totalisations spéciales sont produites en recouvrement des coûts.

Aperçu général de l'enquête

Population cible. L'EPA couvre 98 % de la population canadienne. Les résidents des Territoires du Nord-Ouest, des réserves indiennes et des terres publiques sont exclus du champ de l'enquête, de même que les pensionnaires d'un établissement institutionnel et les membres à temps plein des Forces armées canadiennes, puisque ces deux groupes sont considérés comme n'étant pas sur le marché du travail. L'enquête établit la situation d'activité de tous les membres âgés de 15 ans et plus des ménages sélectionnés.

Taille de l'échantillon. Au moment d'écrire ces lignes, la taille de l'échantillon visée de l'EPA était de 52 350 ménages. Toutefois, ce nombre a varié au fil des ans. À la suite du remaniement des années 1970, l'échantillon mensuel est passé de 35 000 ménages à 55 000 pour répondre à la demande croissante de données provinciales plus fiables et plus détaillées. Les échantillons mensuels ont également été réduits à l'occasion. À la suite de deux diminutions dans les années 1980, l'échantillon comprenait environ 47 000 ménages. En 1989, l'échantillon était augmenté de 16 500 ménages pour atteindre 63 000 ménages. Cette hausse visait à produire de meilleures estimations pour les régions d'assurance-emploi. De ce nombre, il passait à environ 59 000 ménages en 1993. Grâce aux gains d'efficacité découlant du nouveau plan, l'échantillon a été réduit à nouveau en juillet 1995, pour passer à 52 350 ménages. Suite aux modifications législatives apportées en juin 1996, les frontières des régions d'assurance-chômage ont été révisées et les régions ont été renommées régions économiques d'assurance-emploi (REAE).

Renouvellement de l'échantillon. L'EPA est une enquête par panel avec renouvellement de l'échantillon. Celui-ci est divisé en six sous-échantillons représentatifs dont les ménages font partie de l'enquête pendant six mois consécutifs. Un de ces sous-échantillons est remplacé chaque mois ; les ménages qui le constituaient ayant fini de participer à l'enquête. Ainsi, les cinq sixièmes de l'échantillon ne changent pas d'un mois à l'autre, ce qui permet de mesurer les variations mensuelles. Par ailleurs, le fait que les ménages sélectionnés sortent de l'échantillon au bout de six mois, évite un fardeau excessif pour les ménages choisis.

Collecte des données. La collecte des données de l'EPA a lieu pendant la semaine suivant la semaine de référence, qui est normalement celle pendant laquelle tombe le quinzième jour du mois. Statistique Canada emploie pour l'ensemble du pays environ 850 intervieweurs, dont 80 intervieweurs principaux, relevant de cinq bureaux régionaux (BR). Chaque ménage est sondé en personne lors de la première interview, et par téléphone au cours des cinq mois suivants. Les questionnaires sont remplis par l'intervieweur au moyen d'un ordinateur portatif et du mode d'interview assistée par ordinateur (IAO). Chaque jour de la semaine d'enquête, l'intervieweur transmet au BR les données recueillies, qui sont ensuite transmises à Ottawa en vue de leur traitement. La collecte, le traitement et la diffusion des données ont été modernisés et sont réalisés de manière efficace. Statistique Canada arrive ainsi à publier les estimations de l'EPA 13 jours seulement après la fin de la semaine d'interview.

Étant donné l'importance des statistiques produites et la complexité des tâches en cause, chacune des étapes est soumise périodiquement à divers programmes d'évaluation et de contrôle de la qualité.

Remaniement de l'enquête. Après chaque recensement décennal de la population, on remanie le plan de sondage de l'EPA pour tenir compte de l'évolution des caractéristiques de la population et des besoins d'information que cette enquête cherche à satisfaire. Le remaniement effectué après le recensement de 1971 avait été le plus important avant le remaniement actuel. Non seulement le plan d'échantillonnage avait-il été modifié, mais des modifications importantes avaient également été apportées au questionnaire et une nouvelle infrastructure de traitement des données avait été mise en place. Lors du remaniement effectué après le recensement de 1981, on a surtout mis à jour le plan de sondage en tant que tel. Le remaniement actuel est d'envergure, puisqu'il englobe tous les aspects de l'enquête : on a introduit l'interview assistée par ordinateur, le plan de sondage a été modifié, les systèmes de traitement et de diffusion des données ont été complètement mis à jour et le questionnaire a été substantiellement révisé.

CHAPITRE 1 - Introduction

Historique

L'Enquête sur la population active (EPA) a été créée après la Seconde Guerre mondiale pour répondre à un besoin urgent de données fiables et actuelles sur le marché du travail reflétant le passage d'une économie de guerre à une économie de paix. L'enquête a été conçue dans le but de produire des estimations sur l'emploi et le chômage tant à l'échelle régionale que nationale.

Trimestrielle à l'origine en 1945, l'EPA est devenue mensuelle en 1952. En 1960, le Comité interministériel de la statistique du chômage recommandait que l'EPA devienne l'instrument officiel par lequel le chômage au Canada serait mesuré. Après l'adoption de cette recommandation, la demande de données a augmenté, les utilisateurs voulant disposer d'une plus vaste gamme de statistiques sur le marché du travail, notamment des données régionales plus détaillées. L'éventail des estimations produites de cette enquête s'est élargi considérablement au fil des ans, et présente aujourd'hui un portrait détaillé du marché du travail canadien.

Concepts et produits

L'EPA est la seule source officielle d'estimations mensuelles touchant l'emploi total (du travail rémunéré et du travail autonome, du travail à temps plein et du travail à temps partiel) et le chômage. Parmi les principaux indicateurs mensuels publiés, mentionnons, le taux de chômage (défini comme le nombre de chômeurs exprimé en pourcentage de la population active), le taux d'emploi (défini comme le nombre de personnes occupées exprimé en pourcentage de la population totale, soit le nombre de personnes âgées de 15 ans et plus) et le taux d'activité (pourcentage de la population qui est soit occupé ou en chômage). Il s'agit d'une des principales sources d'information sur les caractéristiques individuelles de la population en âge de travailler (notamment l'âge, l'état matrimonial, le niveau d'instruction et la situation familiale).

Les estimations de l'emploi sont ventilées à un niveau détaillé selon, notamment, la branche d'activité, la profession, la durée de l'emploi, le nombre d'heures habituellement travaillées et le nombre d'heures réellement travaillées. Certaines des questions posées permettent d'étudier une grande variété de sujets d'actualité comme le travail à temps partiel non choisi, le cumul d'emplois et l'absentéisme au travail.

Les estimations sur le chômage sont produites par branche d'activité et par profession, ainsi que selon la durée du chômage, le genre de travail recherché et l'activité avant la recherche de travail. Il existe également des données sur l'activité récente sur le marché du travail des personnes actuellement inactives. On trouvera dans Statistique Canada (1998) une description complète du contenu du questionnaire de l'EPA.

Dans le cadre du remaniement de l'EPA, on a introduit un nouveau questionnaire en janvier 1997. Ce questionnaire couvre de nouveaux aspects, notamment les mesures de gains et l'affiliation syndicale. On trouvera dans Sunter et coll. (1995) une description détaillée du processus de remaniement du questionnaire de l'EPA.

En plus des estimations nationales et provinciales, l'EPA produit également des données pour des régions infra-provinciales, comme les régions économiques d'assurance-emploi (REAE) et les régions métropolitaines de recensement (RMR). Ces dernières années, on a également totalisé, à l'aide de techniques spéciales d'estimation, des indicateurs standard du marché du travail pour de petites régions, comme les divisions de recensement (DR) et les centres d'emploi du Canada. Le gouvernement fédéral et les gouvernements provinciaux se servent des données de l'EPA pour la répartition des ressources financières et autres entre les diverses juridictions politiques et administratives.

Les estimations standard de l'EPA paraissent chaque mois dans la publication *Information - population active* (numéro 71-001 PPB au catalogue). On peut également accéder à une variété de données propre au marché du travail par l'intermédiaire de CANSIM, la base de données et le système d'extraction électronique de Statistique Canada. Cette base, qui contient plus de neuf milles séries chronologiques, est mise à jour mensuellement avec les données de l'EPA.

À compter de 1997, la publication trimestrielle *La population active - Mise à jour* (numéro 71-005-XPB au catalogue) examine de manière exhaustive une variété de sujets pertinents à l'analyse du marché du travail. Chaque question est abordée sous un angle particulier, et des

Table des matières

Remerciements

La réalisation et la conception de l'Enquête sur la population active (EPA) exige la participation d'un grand nombre d'intervenants de Statistique Canada. Les divisions ou directions suivantes jouent un rôle de premier plan : la Division des enquêtes-ménages, qui est responsable de la gestion de l'enquête, de la diffusion des données et de la liaison avec les utilisateurs ; la Direction des opérations des enquêtes, qui est chargée des activités sur le terrain et de la saisie et du traitement des données dans les bureaux régionaux ; la Direction de la méthodologie, qui est responsable du plan d'échantillonnage, des méthodes de collecte des données, des méthodes d'estimation et de l'évaluation de la qualité ; la Direction de l'informatique, qui est responsable des services informatiques de traitement des données à Ottawa et dans les bureaux régionaux.

En 1990, on a mis sur pied un Comité de direction du remaniement composé de représentants des secteurs susmentionnés, de même que des services dont les programmes sont liés à l'infrastructure de l'EPA. Ce comité, qui s'est réuni jusqu'en 1996, a orienté et guidé l'élaboration et la mise en oeuvre du remaniement de l'échantillon.

Voici les principaux collaborateurs à la réalisation du présent rapport : Jack Gambino, M.P. Singh, Johane Dufour, Brian Kennedy et John Lindeyer. Ont également apporté leur contribution et prodigué leurs conseils : Doug Drew, Mike Sheridan, Deborah Sunter et Diane Stukel.

Statistique Canada
Division des méthodes des enquêtes auprès des ménages

Méthodologie de l'enquête sur la population active du Canada

J.G. Gambino, M.P. Singh, J. Dufour, B. Kennedy, J. Lindeyer

Publication autorisée par le ministre responsable de Statistique Canada

Décembre 1998

N° 71-526-XPB au catalogue

Périodicité : occasionnelle

ISBN 0-660-60566-X

Ottawa

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Des données sous plusieurs formes

Statistique Canada diffuse les données sous formes diverses. Outre les publications, des totalisations habituelles et spéciales sont offertes. Les données sont disponibles sur Internet, disque compact, disquette, imprimé d'ordinateur, microfiche et microfilm, et bande magnétique. Des cartes et d'autres documents de référence géographiques sont disponibles pour certaines sortes de données. L'accès direct à des données agrégées est possible par le truchement de CANSIM, la base de données ordinolingue et le système d'extraction de Statistique Canada.

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes des enquêtes auprès des ménages, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : (613) 951-9809) ou à l'un des centres de consultation régionaux de Statistique Canada :

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

Vous pouvez également visiter notre site sur le Web : http://www.statcan.ca

Un service d'appel interurbain sans frais est offert **à tous les utilisateurs qui habitent à l'extérieur des zones de communication locale** des centres de consultation régionaux.

Service national de renseignements	**1 800 263-1136**
Service national d'appareils de télécommunications pour les malentendants	**1 800 363-7629**
Numéro pour commander seulement (Canada et États-Unis)	**1 800 267-6677**

Renseignements sur les commandes et les abonnements

Les prix ne comprennent pas les taxes de vente

Le produit n° 71-526-XPB au catalogue paraît occasionnellement en version imprimée standard au coût de 50 $ au Canada. À l'extérieur du Canada, le coût est de 50 $US.

Veuillez commander par la poste, en écrivant à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario) K1A 0T6; par téléphone, en composant le **(613) 951-7277** ou le **1 800 770-1033**; par télécopieur, en composant le **(613) 951-1584** ou le **1 800 889-9734**; ou par Internet, en vous rendant à order@statcan.ca. Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresses. On peut aussi se procurer les produits de Statistique Canada auprès des agents autorisés, dans les librairies et dans les bureaux régionaux de Statistique Canada.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois et dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec le centre de consultation régional de Statistique Canada le plus près de chez vous.

ISBN 0-660-60566-X

9 780660 605661

71-526-XPB 98001

Nº 71-526-XPB au catalogue

Méthodologie de l'enquête sur la population active du Canada

Statistique Canada
Statistics Canada

Canada